MICROFOUNDATIONS OF AGGRESSIVE COMMENTING ON SOCIAL MEDIA WITHIN A SOCIOLOGICAL MULTILEVEL PERSPECTIVE

Thesis (cumulative thesis)

presented to the Faculty of Arts and Social Sciences

of the University of Zurich

for the degree of Doctor of Philosophy

by Lea Stahel

Accepted in the spring semester 2018

on the recommendation of the doctoral committee:

Prof. Dr. Katja Rost (main supervisor)

Prof. Dr. Sophie Mützel

Zurich, 2018

Table of contents

List of figures

List of tables

List of documents

1. Synopsis: Microfoundations of aggressive commenting on social media within a sociological multilevel perspective

1.1. Summary

Social media have democratized the public expression of opinions not only for those who debate constructively but also for those who vent their anger and attack and silence others. In recent years, aggressive online commenting in social media has become highly visible, common, and socially relevant. Politicians are threatened on Facebook, organizations are showered in collective outrage, and journalists are defamed by their audiences. This form of online aggression occurs predominantly in social-political settings, is highly public, is collective, and targets public actors. Explaining and empirically exploring it is the focus of this dissertation. Compared to the three studies incorporated in this dissertation, the value added by this synopsis is generated by embedding the three studies in an overarching metatheoretical framework and discussing their contributions to it.

The first section outlines the current state of scientific knowledge about aggressive online commenting. It shortly reviews this concept and its historical development and then outlines the multidisciplinary explanations so far proposed for it. This section closes by identifying one major theoretical and one methodological gap in the literature.

The second section outlines the sociological multilevel perspective as a metatheoretical framework that connects all three studies. In addition, it introduces the subtheoretical approaches of social norms, legitimacy, and social identity applied in the studies, and how they can be located within this meta-theoretical framework.

The third section provides an overview of the three studies and outlines their broader scientific contributions. The first study (Rost, Stahel and Frey 2016) introduces a social norm explanation of online aggression. It explains why non-anonymous individuals in norm-enforcing settings are more aggressive than anonymous individuals. This is explored in 532,197 comments of 1,612 online petitions. The second study (Stahel and Rost 2018) elucidates the microfoundations of legitimacy construction. It predicts that online users who judge organizational character, procedures, structures, and outcomes as morally illegitimate are motivated to attack organizations on social media. This is confirmed through manual and automated content analysis of 45,997 firestorm comments. The third study (Stahel 2018) asks which journalists are particularly frequently attacked by their audiences and why. A survey of

530 journalists in Switzerland finds that journalists with higher potential to threaten the social identity and power of social groups are more frequently attacked.

1.2. Introducing online aggression

1.2.1. Concept and historical transformation

Aggressive online commenting in social media (termed *online aggression* in the following) is a multifaceted concept that has no uniform definition. According to Jane (2015: 83), online aggression "takes countless forms in countless domains involving countless participants". This makes its exploration challenging conceptually, methodologically, and epistemologically (see also O'Sullivan and Flanagin 2003). Not only individuals but also organizations, institutions, ideas, and groups can be attacked in a variety of online contexts based on numerous characteristics such as ethnicity, religion, gender, age, disability, political conviction, and status. Consequently and unsurprisingly, no universally preferred term or definition exists for online aggression (Jane 2015, Ksiazek, Peer and Zivic 2015). In this dissertation project, online aggression is defined as expressing threats, insults, hostilities, derogatory statements, and vulgarities through online comments via social media. Social media are all web-based means for individuals to connect and interact with content and other users and to generate and distribute content on platforms such as social network sites, blogs, news commentary sections, and email (see Treem, Dailey, Pierce et al. 2016). The definition of online aggression proposed here covers the central elements of the strongly overlapping definitions presented in the current literature.

Although online aggression has been researched since the beginning of the Internet, it has risen to scientific and societal prominence only in recent years. The first studies on online aggression appeared over 30 years ago (see Kiesler, Seigel and McGuire 1984). For a long time, online aggression was a niche phenomenon and primarily observed within and between similar-status members of Internet subgroups (Kayany 1998) and adolescents in school contexts (Bauman 2013). In recent years, though, it has changed. Today, incivility in public discourse attracts more concern as it "has more outlets, can be highly public, and travels and spreads faster; its impact can be greater, and strategies for dealing with it are still being tested" (Meltzer 2015: 86). These concerns are also reflected in Jane's (2015: 67) observation that online aggression has developed into an "acute social problem", as it is increasingly common and even the norm rather than the exception in various online contexts. Although recent online aggression cannot be strictly distinguished from the more traditional forms, the more recent aggression seems more often to share the following characteristics. It occurs in highly public, social-political

settings rather than private niche contexts or within specific social groups. It targets public, often higher-status actors such as politicians, corporations, NGOs, and journalists. It is also collective: this means that it aims at collective outcomes, whether perpetrated alone, as is commonly observed in digital environments, or performed by more than one person. Publicity and collectivity both contrast with traditionally interpersonal interactions between same-status users. To the broader public, this recent online aggression is typically visible through the uppermost phenomena of the iceberg, for example online firestorms. They describe, large volatile amounts of criticism, insulting comments, and swearwords in social media, which are often fueled by media attention (Pfeffer, Zorbach and Carley 2014). Increasingly, scientific findings show how online aggression harms the targeted individuals' well-being, social harmony, and cohesion. For example, victims of online aggression are more likely to become depressive, suicidal, and cyberbullies themselves (Dooley, Pyzalski and Cross 2009, Bauman 2013). In addition, exposure to online aggression can make the reading audience think and act in a more hostile manner (Hsueh, Yogeeswaran and Malinen 2015, Rösner, Winter and Krämer 2016, Kwon and Gruzd 2017, Masullo Chen and Lu 2017), distrust news media, science, politics, and people in general (Borah 2013, Näsi, Räsänen, Hawdon et al. 2015, Nauroth, Gollwitzer, Bender et al. 2015, Anderson, Yeo, Brossard et al. 2016, von Sikorski and Hänelt 2016), and harden and polarize their opinions (Anderson, Brossard, Scheufele et al. 2014, Borah 2014). It is no surprise that effective responses to online aggression are widely sought by government actors, social media corporations, and civil society. Explaining what drives recent aggression is thus a highly relevant first step addressed in this dissertation.

1.2.2. Existing multidisciplinary explanations

Although recent aggression is explained here, this dissertation builds on existing theories and empirical evidence on online aggression to address both more traditional and more recent forms. The scholarly field presents a kaleidoscope of theories to explain the complex and multifaceted nature of online aggression (see section 1.2.1.). It is multidisciplinary with a fragmented scholarship, predominantly from information science, political science, psychology, education, communications, journalism, marketing, and business. A first stream conceptualizes and addresses online aggression as a more general phenomenon explored across disciplinary borders. This includes literature on flaming as a general term for online aggression (O'Sullivan and Flanagin 2003), trolling for provoking for fun (Hardaker 2010), and online firestorms (Pfeffer et al. 2014, Poerksen 2018). Another, more distinct stream more clearly delineates aggression. Typically, this includes the massive volume of psychological literature on cyberbullying, which is limited predominantly to minors and college-age adolescents in

educational contexts (Slonje and Smith 2008, Smith, Mahdavi, Carvalho et al. 2008, Bauman 2013, Mehari, Farrell and Le 2014). In addition, political and legal studies on online hate speech and online incivility conceptualize online aggression as predominantly democracy-threatening behavior (Anderson et al. 2014, Rowe 2015). Other examples include economy-related studies on cyber-smearing, online dysfunctional customer behavior, and corporate harassment. The slightly pejorative connotation of these terms indicates its conceptualization as a threat to economic actors (Tuzovic 2010, Workman 2012, Kim, Wang, Maslowska et al. 2016).

This fast-growing multidisciplinary literature on online aggression has proposed a variety of reasons why individuals are aggressive on social media. In the following, I identify and review the two most established explanatory meta-approaches, since they lead us to the central gaps identified in this dissertation and justify the theoretical approach it proposes. I name these meta-approaches *individual-psychological explanations* and *immediate digital- and social-context explanations*.

The individual-psychological explanations root online aggression in stable psychological traits (i.e. aggressors as antisocial individuals) and fluid individual emotions and goals (i.e. aggression as a tool to vent or to instrumentalize). The theory on traits proposes that each individual has a unique personality and that associated traits motivate online aggression. For example, online aggressors score higher than non-aggressors in narcissism, psychopathy, and Machiavellianism (Buckels, Trapnell and Paulhus 2014, Pabian, De Backer and Vandebosch 2015, Koban, Stein, Eckhardt et al. 2018), may lack empathy (Steffgen, König, Pfetsch et al. 2011, Sticca, Ruggieri, Alsaker et al. 2013) and self-control (Alonzo and Aiken 2004, Peterson and Densley 2017), and are also more shy and prone to depression (Bauman 2013, Bonanno and Hymel 2013). Individuals who are aggressive on social media were also proposed to be driven by emotions such as negative mood (Cheng, Bernstein, Danescu-Niculescu-Mizil et al. 2017) and anger (Johnson, Cooper and Chin 2009, Stephens, Trawley and Ohtsuka 2016). People also use aggression instrumentally to seek thrills and fun, to draw attention to social injustice (Erjavec and Kovačič 2012), and to gain social standing (e.g. Kasumovic and Kuznekoff 2015, Wright 2017). Overall, these explanations are individual-focused and often apply emotional-motivational theories. The cyberbullying literature in particular focuses on describing the conceptual and typological development of online aggressors while commonly lacking theoretical explanation (Mehari et al. 2014: 2).

The immediate digital and contextual explanations root online aggression in technology (i.e. technologies motivates aggression) and in immediate contextual factors (i.e. certain online

surroundings motivate aggression). In the 1980s and 1990s, social psychologists explored whether online aggression results from computers or from social online contexts; the final answer was both (Jane 2015). Advocates of the reduced-cues approach (Kiesler et al. 1984) argued that online environments cause people to be toxically disinhibited (Suler 2004) because communication lacks social-context cues and is anonymous, invisible, and asynchronous. However, opponents argued that aggression is not an inevitable byproduct of technology. Instead, online users conform to surrounding social cues on platforms. This social context explanation has been broadly supported (drawing e.g. on social learning theories or on the social identity model of deindividuation effects model by Reicher, Spears and Postmes 1995). For example, online users are more aggressive if they perceive aggression to be socially acceptable on the platform (Moor, Heuvelman and Verleur 2010, Hmielowski, Hutchens and Cicchirillo 2014), if they are attacked by other users (Hutchens, Cicchirillo and Hmielowski 2015, Masullo Chen and Lu 2017, Quintana-Orts and Rey 2018) if they are not sanctioned by others (Xu, Xu and Li 2016, Álvarez-Benjumea and Winter 2018), if certain social identity processes are activated (Nauroth et al. 2015, Rains, Kenski, Coe et al. 2017), and if platforms set up moderation strategies that fail to prevent aggression (Ruiz, Domingo, Micó et al. 2011, Stroud, Scacco, Muddiman et al. 2014, Ksiazek 2015). These explanations incorporate the social context, though they commonly limit it to the immediate online environment.

Both meta-explanatory approaches draw on a variety of data sources and designs. Surveys that collect either self-reported intended or actual aggression predominate, particularly in exploring adolescences' and students' cyberbullying. Experiments are also often conducted. Common samples include MTurk participants, gamers, and social media users. Further, content analysis of reader comments on social media platforms is common, often in reference to a specific topic. Comment data sets commonly include a few hundred online comments. A few studies have drawn on several hundred thousand comments (see Stephens et al. 2016, Kwon and Cho 2017, Vargo and Hopp 2017). Firestorms have been discussed theoretically as prototypical examples of recent online aggression (Pfeffer et al. 2014) and explored in case studies (Salek 2015), simulations (Hauser, Hautz, Hutter et al. 2017), and experiments (Johnen, Jungblut and Ziegele 2017). Content analyses of firestorm comments exist but are scarce (see Prinzing 2015, Franke, Keinz, Taudes et al. 2017).

1.2.3. Gaps

Based on the literature review, one major theoretical gap and one methodological gap have been identified. Theoretical conceptualization and explanation of more recent forms of online aggression are scarce, particularly in regard to their societal dimension. Today, online

aggression does not only involve pupils privately cyberbullying their classmates or online gamers mobbing fellow gamers to gain status. Instead, it incorporates a societal dimension as it seems to span social, geographical, and hierarchical structures and accordingly produces societal effects (see section 1.2.1.). Nevertheless, most current explanations of online aggression are limited to individual and immediate digital environment factors (section 1.1.3.) and ignore its embeddedness in the broader societal context, including any reciprocal links between aggressors and society. Scholars have begun to acknowledge this gap. Accordingly, Jane (2015: 66) notes that "scholarship in many disciplines has missed the growing power and prevalence of hostility on the internet". Similarly, Peterson and Densley (2017: 197) call other scholars to address "the interaction between micro, meso, and macro levels of explanation" to overcome the current lack of "continuity and coherence" in online aggression research. Suggested broader explanations also coincide with DiMaggio, Hargittai, Neuman et al.'s (2001: 329) more general call for "explanatory models […] of Internet use […] that tie behavior directly to social and institutional context". Recently published studies have followed these calls and adopted broader social approaches (see Nauroth et al. 2015, Einwiller, Viererbl and Himmelreich 2017, Johnen et al. 2017, Rains et al. 2017, Vargo and Hopp 2017). However, they remain scarce.

The rarely explored societal dimension opens associated gaps, especially the limited focus on adult producers of aggression and public targets. The lion's share of research into online aggression has accounted for a school problem involving young people. However, "adult cyberbullying remains unstudied, even though it is a pressing social problem and a dark side of the Internet" (Simons 2015, Lowry, Zhang, Wang et al. 2016: 981). It is plausible that recent social-political online aggression is more likely driven by adult aggressors. Further, although recent aggression predominantly targets public high-status actors, it has not been explicitly theorized which public actors are more likely to be targeted than others or why. This gap cannot be systematically answered by studies of youth exposed to and targeted by online hate (Costello, Hawdon and Ratliff 2017, Hawdon, Oksanen and Räsänen 2017) or descriptive or case studies of targeted corporations (Pace, Balboni and Gistri 2017), journalists (Löfgren Nilsson and Örnebring 2016), or social groups generally (Silva, Mondal, Correa et al. 2016) such as female bloggers (Hardaker and McGlashan 2016).

To conclude, there is a need to address recent online aggression and to embed it in a socially comprehensive perspective. Such an approach more effectively accommodates the nature of recent online aggression, works out its societal relevance, and predicts and explains aggressive acts and who they target.

Addressing the methodological gap requires behavioral data sets of recent online aggression that are larger, broader, comparable, and include control cases. While in the last two years a few studies using large sets of comments involving aggression have emerged (e.g. Hewett, Rand, Rust et al. 2016), they are mostly limited to specific topics and thus likely reflect particular social groups. To draw generalizing conclusions about why online aggression occurs, data is needed on more diverse groups of commentators (Hutchens et al. 2015, Runions, Bak and Shaw 2017, Koban et al. 2018) and on broader societal topics (Sullivan 2014, Jane 2015). Simultaneously, data should be comparable. This can occur if data share the same platform. This avoids bias arising from differing platform architectures and participation conditions such as moderation strategies. Also, scholars studying firestorms commonly select and study protests labelled as firestorms by news media and high-profile single actors that attracted firestorms (also criticized by Sullivan 2014). However, selecting for outcome bears the risk of ideological bias and decreases generalizability. Instead, selection should be guided by objective criteria that consider the actual content and provide non-aggressive control cases. In addition, aggressive comments should be comparable with each other and with non-aggressive comments. For instance, online protests that turn into firestorms should be compared to others that remain non-aggressive. These methodological gaps also apply when examining who is targeted by aggression. Here, essay-like, qualitative, and descriptive studies currently predominate. Thus, entities within representative samples of professional, organizational, and other groups should be compared in statistical, multivariate models according to how frequently they are attacked. This gap reflects a broad call in literature for behavioral datasets of online aggression and of attacked targets that are larger and from a broader societal range but are comparable and not selected for outcome. Such a methodological approach increases the external validity, generalizability, and reproducibility of results.

1.3. Metatheoretical approach: Sociological multilevel explanation

To address the theoretical gap, the following section will present the metatheoretical, sociological multilevel approach that will link all the dissertation studies. The section also introduces the three themes underlying the studies: social norms, legitimacy, and social identity. They all traditionally underlie multilevel thinking. The idea of this dissertation has never been to strictly apply the meta-theoretical framework with all its analytical steps. Rather, this framework is used as an encompassing background perspective that guides the studies, helping identify research questions, selecting sub-theories and variables, and interpreting, and drawing conclusions (for a discussion of meta-theories, see Wagner and Berger 1985). In addition, the metatheory helps to locate this dissertation explicitly within the online aggression literature and

especially within sociology. Overall, this metatheoretical approach is suited to explaining aggressive behavior as it focuses on explaining micro behaviors, while acknowledging the social-structural macro context.

Accordingly, the proposed sociological multilevel approach uses an explanatory logic that is shared most popularly by, for example, the micro-macro diagram by Coleman (1990), contemporary analytical sociology (Hedström and Bearman 2009, Hedström and Ylikoski 2014), explanatory sociology (Wippler and Lindenberg 1987, Maurer and Schmid 2010), rational choice approaches in social science (Diekmann and Voss 2004), and behavioral and experimental game theory (Camerer 2011). Its roots can be traced back to Scottish moralists such as David Hume and Adam Smith (Raub, Buskens and Van Assen 2011) and to social exchange theory (Homans 1958, Blau 1964).

This approach conceptualizes sociology as an explanatory discipline that strives to link individual actions to the structural level (Raub et al. 2011). The multilevel approach takes up sociology's basic question of how social phenomena, regularities, and facts such as the preservation of social order can be explained by the reciprocal relationships between individuals and society. In essence, it explains macro-level facts by micro-level, individual mechanisms: individuals' actions and the relations that link actors to one another (Hedström and Bearman 2009). This enables an understanding of why the observed outcomes were produced. This idea commonly underlies methodological individualism (Weber 1978 [1922]) and structural individualism, according to which "all social facts, their structure and change, are in principle explicable in terms of individuals, their properties, actions, and relations to one another" (Hedström and Bearman 2009: 8). The linking of micro and macro levels in this approach thus avoids both purely micro explanations and purely structural macro holism such as that practiced by Durkheim and Parsons (Raub et al. 2011). In addition, it goes beyond mere sociological description, typologies, and hypotheses of orientation such as the risk society. Although these are important predecessors of explanations, they provide no information on why things are the way they are. Paradigmatic examples of this micro–macro thinking is the production of collective goods (Olson 1965), Schelling's (2006) models of segregation, collective action, and the production of status hierarchies.

The micro–macro link is a central idea in the multilevel approach. It is thus shortly described here. I draw on Coleman's (1990) micro–macro diagram as it is the standard tool for representing micro–macro models and for its "simplicity and intuitive appeal" (Raub and Voss 2017: 2). Coleman (1990) distinguishes the macro level, which refers to collective properties

of the social situation and social system such as a society, firm, and social group, from the micro level, which refers to cognitive properties and purposive actions of individual actors (Raub and Voss 2017). The model is divided into three different steps.

The first step describes how the macro level affects the micro level. Whenever people decide whether to vote, to move to a new neighborhood, to buy the latest best-seller, and to express an opinion, they unconsciously or consciously consider macro properties (Hedström and Bearman 2009). Macro properties include typical actions and beliefs among members of the collectivity and aggregate patterns that characterize the collectivity, such as distributions of inequality, norms, positions, networks, and institutions (Hedström and Bearman 2009). Individuals identify and evaluate these properties and their associated conditions, incentives, alternatives for action, objective restrictions for action, and consequences to be expected. This step thus reconstructs the social situation as perceived by the actor, including potential bias, as actors are not assumed to be fully informed (Wippler and Lindenberg 1987).

The second step explains why individuals act: how individuals' micro evaluations, goals, and expectations affect which action individuals select. It involves a micro theory of individual behavior that may be inspired by other disciplines such as psychology (Maurer and Schmid 2010). Social sciences often assume actors to be boundedly rational; individuals act rationally and are predominantly self-interested by and large, so they can be motivated by both material and non-material incentives (Diekmann and Voss 2008, Raub et al. 2011). For example, individuals may expect an action to maximize or optimize a utility value, an emotional value, or a degree of appropriateness (Maurer and Schmid 2010). A main task of this step is to identify the microfoundational social cogs and wheels that explain action (Hedström and Bearman 2009). These may include norms when centering on others' actions but also individual-centered causes such as beliefs, altruistic motives, and heuristics (Hedström and Bearman 2009, Maurer and Schmid 2010).

The third step addresses how individuals' micro-level actions generate macro-level social phenomena and collective outcomes. Diverse transformation rules can be applied (Wippler and Lindenberg 1987), ranging from the mere addition of actions such as collecting signatures to complex diffusion models such as threshold models. This step is a common challenge and goes beyond the micro theory of action (Maurer and Schmid 2010). Generally, surveys are particularly suited to empirically revealing multilevel models, beside other methods such as agent-based computer simulations.

The following reviews the three themes underlying the three studies. As they are traditionally explained by micro-macro processes, their embeddedness in multilevel thinking is discussed. They suit the overarching research question of this dissertation well thanks to their combination of macrostructural foundation with primary focus on micro processes.

1.3.1. Social norms

Norms is a concept that has traditionally been theorized from a micro-macro perspective. Fundamentally, a norm is "a statement that something ought or ought not to be the case" (Opp 2002). In sociology, norms are commonly understood as institutionalized expectations of behavior, such as roles (Dahrendorf 1985/2010), or as regularity of behavior (Diekmann and Voss 2008, Tutić, Zschache and Voss 2015). Social norms are those shared by other people and sustained by sanctions: the approval and disapproval of third parties (Homans 1950, Elster 1989, Bendor and Swistak 2001). Norms have been variously conceptualized as a macro-level property, as part of the collective consciousness, as a culture-specific social fact (Durkheim 1957) or, in line with methodological individualism, as a prototypical macro-level product arising from individual actions (Elster 1989, Coleman 1990: 244). Norms can also exist at the micro level if they cause individuals to perceive norms, to follow or violate them, and to enforce conformity (Elster 1989, Coleman 1990). Modern society is characterized by particularist, co-existing political, religious, and philosophical norms (Popitz 1980), each with their own standards of norm-conforming behavior (Weber 1904).

Norms permeate societies. They prescribe and proscribe how and how not to dress, eat, queue, distribute, retaliate, and cooperate (Elster 1989). How norms emerge, affect society, and particularly how they are enforced is studied by philosophers, economists, legal experts, and sociologists, particularly those using an individualistic theory of bounded rational choice (Tutić et al. 2015). People who enforce norms through sanctions play a central role in how social behavior, social cohesion, and social order is generated (Diekmann and Voss 2008, Scherr 2013).

How norms can be enforced by rationally thinking actors is a classic social dilemma. Often, individual micro rationality, a state in which self-interested people prefer to free-ride, leads to collectively suboptimal outcomes on the macro level in which the norm is not enforced (see second order collective good problem; Diekmann and Voss 2008). Basically, if actions have negative effects for other individuals, termed externalities, such as environmental pollution, norms to restrict and forbid such actions may be introduced (Opp 2002, Elster 2015) and enforced such as by naming, shaming, and blaming norm violators. Costly sanctions by many

effectively produce a bad reputation through gossiping, ostracizing, and adjusting exchange relationships with norm violators (Coleman 1990). This is facilitated by digital communication technologies, which enable geographically scattered individuals with narrow common interests to act collectively (Elster 1989). However, applying informal sanctions is costly. It takes time, sanctioners risk opportunity costs, and norm violators may retaliate. Simultaneously, sanctions benefit all members of the group, including those freeriding on anyone who volunteers as social police (Olson 1965). The dilemma can be solved if norm enforcement is cheap, if sanctioners are materially, emotionally, and socially rewarded, and if they are intrinsically motivated (Diekmann and Preisendörfer 1992, Fischbacher, Fehr and Gächter 2001, Opp 2002). Overall, norm violations and subsequent scandalizations thus perform the indispensable social function of strengthening or changing the macro validity of social norms (Durkheim 1977).

1.3.2. Moral legitimacy

Legitimacy is a second concept that has been theorized from a micro–macro perspective. Legitimacy is central to sociology and relates to other socially evaluative concepts such as status and reputation (Deephouse and Suchman 2008). Weber (1978 [1922]) and Parsons (1960) are commonly credited with introducing legitimacy into sociological theory. Weber analyzed the legitimacy of traditional, charismatic, and legal authority types that are needed to produce stable social orders. In addition, he argued that legitimacy occurs through a collective construction of social reality. Social order is legitimate and a social fact if it conforms with widely shared norms, values, and beliefs. Building on this, Parson viewed legitimacy as the congruence of an organization with social laws, norms, and values. Accordingly, organizational legitimacy, as researched in organizational studies and new institutional theory, has emerged as a central stream in legitimacy research (Deephouse and Suchman 2008).

Suchman (1995: 574) offered a broad-based definition of legitimacy as "a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions". A focal interest is how organizations gain, maintain, and repair legitimacy (Suchman 1995); it is not enough to have material and technical resources to survive economically (Deephouse and Suchman 2008). Suchman (1995) suggested conceptual dimensions of legitimacy according to which audiences such as state regulators, society at large, and media assess organizational legitimacy. One dimension is moral legitimacy. This describes whether the consequences, procedures, persons, and structures of organizational activities are perceived as right or wrong by audiences. Audience judgements influence legitimation and delegitimation, the processes by which the legitimacy of an entity changes over time (Deephouse and Suchman 2008).

Legitimacy construction has been conceptualized as a micro–macro issue. Following Deephouse and Suchman's (2008) suggestion that legitimation should be examined at multiple levels, Bitektine and Haack (2015) proposed a multilevel theory of legitimacy construction focused on microfoundations. This was further developed in the legitimacy-as-perception approach by Suddaby, Bitektine and Haack (2017). It conceives legitimacy as a collective process that is mediated by individual perceptions and their associated socio-cognitive processes, which guide behaviors. On the macro level, legitimacy is a socially shared opinion about the appropriateness of an organization, termed its validity. On the micro level, individuals have certain beliefs about their own and others' perceptions of the legitimacy of an organization, termed validity and propriety beliefs. To form judgements, which guide actions, individuals often use socio-cognitive heuristics (Kahneman and Frederick 2002, Gigerenzer 2008). Aggregating micro perceptions and actions can support or challenge organizational legitimacy at the macro level.

1.3.3. Social identity

Social identity is a concept that has been traditionally theorized from a micro-macro perspective. The concept of identity spans many social science disciplines (Stryker and Burke 2000). In the sociological and social psychological traditions, identity theory and social identity theory are most central. Both theorize the social nature of self as constituted by society and social structures. The focal interest lies in how identities are internalized, become part of the self, and influence social behavior. Identity theory can be traced back to Mead (1934) and Blumer (1969) and stems from the associated micro-sociological approach of symbolic interactionism (Stryker 1980). Identity theory posits that the impact of society on behaviors is mediated by role identities. The self reflects the wider social structure insofar as it is constituted by multiple identities derived from role positions, the expectations attached to positions that individuals hold in social networks. If people internalize role expectations, they identify. Society and social order is thus enabled through people communicating through their roles (Mead 1934).

The subsequent social identity theory (Tajfel and Turner 1979, Tajfel and Turner 1986) originated in social psychology. It bears various similarities to identity theory (Hogg, Terry and White 1995, Stets and Burke 2000) but primarily explains group processes and intergroup relations such as collective action and discrimination. Social identity describes individuals' awareness of their membership in a social group or groups, including the value and emotions attached to being members (Tajfel and Turner 1986). If individuals self-categorize into groups, their perception, feelings, and actions become depersonalized, and they conform to the self-

defining in-group prototype. If social identities are threatened, self-defense mechanisms may motivate members to disparage the threatening source. As with identity theory, society is here assumed to influence behavior via the mediation of social identity and associated self-categorization processes. The concomitant advantage of social identity lies in its systematic elaboration of the psychological socio-cognitive processes underlying the operation of identity, with the sociological level acknowledged by socio-historical intergroup relations (Hogg et al. 1995). The politicized social identity perspective (see Simon and Klandermans 2001) reflects this multilevel idea particularly well.

This, social identity theory attends to both macro and micro processes (Stets and Burke 2000). As a social-psychological identity theory, it forms part of the structural social psychology approach. This approach was proposed by Lawler, Ridgeway and Markovsky (1993) in response to Coleman's (1990) emphasis of micro-processes in micro–macro analysis. Lawler et al. (1993: 269) argued that social psychological aspects connecting individuals to social structures should be considered in multilevel theories. These may be few and simple aspects, as in rational choice theories, or more complex, as in theories of the self and identity. Social-psychological work on power, status, roles, and identity offered itself to such multilevel theorizing due to its concern with the emergence and effects of social structure (Lawler et al. 1993).

1.4. Overview of studies

The following section provides a short overview of the three studies.

The first study, titled "Digital Social Norm Enforcement: Online Firestorms in Social Media" (Rost et al. 2016), has two goals. First, it introduces social norm theory to conceptualize online aggression in social-political online settings. Second, it uses this social norm explanation of online aggression to challenge the popular assumption that online aggression can be predominantly explained by anonymity. The study predicts that non-anonymous individuals are more aggressive than anonymous individuals, particularly if selective incentives are present and if aggressors are intrinsically motivated. Predictions are confirmed in a multivariate analysis of 532,197 comments on 1,612 online petitions.

The second study, titled "Legitimacy perceptions in online firestorms" (Stahel and Rost 2018), focuses on aggressive acts in online firestorms against organizations. The study integrates the concept of moral heuristics (Sunstein 2005) into the legitimacy-as-perception approach to elucidate the micro-level cognitive processes leading to aggressive actions. It predicts that judging organizational character, procedures, structures, and outcomes as morally illegitimate

motivates commentators to attack organizations in social media. This is empirically confirmed in a lexicon-based, manual, and automated content analysis and subsequent multivariate regression of 45,997 comments in a firestorm against a German music rights organization.

The third study, titled "*Dirty journalists, all liars!* - A social identity explanation for why journalists are attacked by audiences" (Stahel 2018), focuses on why some journalists seem to be more frequently aggressively targeted by their audience through digital and analogue channels. It draws on five central conditions that a politicized social identity approach suggests increase threats to groups' social identity and power. Journalists should be more frequently attacked if they are evaluative, publish on political topics, have a local focus, are powerful, and belong to similar outgroups: other groups that are hard to distinguish from the ingroup. This is because their potential to threaten the social identity and power of groups is greater under these conditions; they are more likely to mobilize group members with a pre-existing tendency to feel threatened. A Swiss online survey and multivariate analysis of 530 journalists confirms all hypotheses.

1.5. Contributions

This section points out the major contributions of the studies and the overall dissertation based on the research gaps, the proposed meta-theoretical framework and its sub-themes (section 1.2.3.).

1.5.1. Theorizing recent forms of online aggression

By exploring more recent online aggression (see section 1.2.1.), all studies contribute knowledge to this scarcely explored phenomenon. All studies thus enlighten aggressive commenting in societally relevant, social-political contexts. This includes online protests about public issues such as internet security, misbehavior of politicians, public spending, and animal protection (Rost et al. 2016), against organizational reforms (Stahel and Rost 2018), and the news media context (Stahel 2018). The aggression explored is collective even if it occurs alone, since the aggressors act on behalf of their own or of other social groups (Rost et al. 2016, Stahel 2018) or in numerically large firestorms (Stahel and Rost 2018). Aggressive commentators in all studies reflect a digital civil society in which particular interest groups become situationally mobilized, such as those concerned with the music service in Stahel and Rost (2018). Aggression occurs mostly on highly public social media, such as on comment sections of online petitions (Rost et al. 2016, Stahel and Rost 2018) and on comment sections of news reported by journalists (Stahel 2018). Finally, targets are primarily public actors and mainly of higher

status, such as authorities, persons, and institutions (Rost et al. 2016), organizations (Stahel and Rost 2018), and journalists as the highly public "fourth estate" (Stahel 2018).

This more recent aggression differs from traditionally occurring and explored aggression in more private, interpersonal contexts between individuals of similar status or known personally to each other (see section 1.2.1.). Naturally, the proposed distinction between older and recent forms should not obscure the fact that much aggression occurs somewhere between or can mutate from one form to another. For example, private cybermobbing may turn into country-wide witch hunts with news media involved.

1.5.2. Microfoundations within a sociological multilevel explanation

All three studies implicitly use a metatheoretical, sociologically multilevel framework in their conceptualization and explanation of online aggression. This framework basically assumes that individuals' embeddedness in society and macro structures contributes to explaining why individuals comment aggressively on social media. This broader sociological perspective focused on microfoundations (Coleman 1990, Diekmann and Voss 2004, Maurer and Schmid 2010, Hedström and Ylikoski 2014) goes beyond most former, narrower approaches to explaining online aggression. Those mainly ignore the macro context and explain online aggression through the individual's psychology or by the immediate digital environment and the presence of social cues on platforms (see section 1.2.1.). The proposed perspective, thereby, is in line with the more general idea of a sousveillance society in which everybody watches everybody through monitoring, capturing, and disseminating information, allowing the denunciation of abuse by the powerful but also enabling intrusive mob vigilantism (Dennis 2008, Ganascia 2010). To draw the proposed, more complete, and socially integrative picture of online aggression, this dissertation project required on the collation and development of the scattered academic literature on online aggression and social behavior. This interdisciplinary integration is reflected by the diverse theoretical approaches selected in the three studies.

The following section discusses how each of the three studies contributes an empirical application of the meta-theoretical framework to online aggression. Aggression is motivated by considerations associated with social norms (Rost et al. 2016), moral legitimacy (Stahel and Rost 2018), and social identities (Stahel 2018). Although the studies neither theoretically nor empirically reflect all aspects of the micro-macro model, they together represent an initial coordinated attempt to open up broader sociological thinking about online aggression.

Study 1

The first study (Rost et al. 2016) establishes the micro–macro link by conceptualizing micro online aggression as an act to enforce macro social norms. In cases of norm violations, online aggressors publicly express disapproval to secure public goods, such as the honesty of politicians. Incremental sanctions (Coleman 1990) in online firestorms are a typical example of such sanctioning behavior. The predominant macro-level properties in this study are thus social norms produced by micro actors, in line with Coleman (1990). The study applies the public good dilemma: rational users will not attack to secure the public good, so the norm is unlikely to be enforced. However, micro aggression can be motivated by low costs through communication technologies, selective incentives made salient by a scandal or a controversy (a macro-level property), and intrinsic motivation (a micro-level property). Micro aggressions ultimately help to secure the public good on the macro level. Most importantly, the results show that non-anonymity (also a macro-level property in the strict sense) motivates aggression because, in this norm-enforcing context, individuals stand up for higher-order moral ideals and principles. Here, anonymous commenting would be a costly, wasteful behavior, as aggressive sanctions are less credible, create less awareness and less support, and offer few benefits from ideologically like-minded networks.

This study enriches literature on online aggression by introducing a novel theoretical approach. In addition, it contributes to the fundamental, widely debated question at the interface of economic and sociological literature about how norms are enforced (Homans 1950, Elster 1989, Bendor and Swistak 2001), and, specifically, how the classic public goods dilemma can be resolved (Olson 1965, Diekmann and Preisendörfer 1992, Opp 2002, Diekmann and Voss 2008). Addressing this question by applying it to aggressive sanctions through digital technologies is innovative as it provides initial knowledge on norm-shaping processes and the societal function (Durkheim 1977) of online aggression in increasingly digitalized societies.

More specifically, the social norm perspective on online aggression entails novel variables - selective incentives and intrinsic motivation - and novel predictions, about non-anonymity. It thus challenges popular assumptions about online anonymity as the main cause of aggression, helps resolve former inconsistent empirical findings on the effect of anonymity on aggression, and provides the argument underlying why non-anonymity motivates aggression in recent online aggression. Thereby, the resulting micro-level evidence disconfirms the popular macro-level assumption that abolition of online anonymity could prevent online firestorms.

Study 2

The second study (Stahel and Rost 2018) connects micro to macro by conceptualizing micro aggressive acts in firestorms as a tool for shaping macro organizational legitimacy. The legitimacy of organizations is constructed, among other things, by online users in social media. The macro level is acknowledged by assuming a socially shared macro opinion about the appropriateness, or legitimacy, of an organization; it affects individuals but is also affected by them. At the micro level, individuals have certain beliefs about whether they think of the organization as legitimate. In construing beliefs, they may use moral heuristics. The resulting beliefs determine whether online users comment aggressively or not. Results show that individuals comment more aggressively if they judge the organization's characters, procedures, structures, and outcomes to be morally illegitimate. This is because moral heuristics introduce ideological justifications and associated emotions that reduce individuals' external and internal costs associated with aggressive acts.

This study thus relates to the well-known concept of legitimacy in sociological and organizational tradition and how organizations gain and maintain legitimacy (Parsons 1960, Weber 1978 [1922], Suchman 1995, Deephouse and Suchman 2008). Specifically, it contributes to understanding the cognitive microprocesses underlying organizational legitimacy construction, an understanding that is still in its infancy (Bitektine and Haack 2015, Etter, Colleoni, Illia et al. 2017). It does so by focusing on social media audiences that challenge organizational legitimacy. It contributes a broader, multi-level perspective by analyzing online aggression innovatively from the explicit multilevel legitimacy-as-perception perspective (Suddaby et al. 2017). The proposed integration of moral heuristics (Sunstein 2005) into the legitimacy-as-perception approach and its effect on online aggression in social media extends and innovatively connects online aggression literature, sociological-organizational, and cognitive psychological literature. Overall, it improves understanding of how legitimacy is constructed in a society in which social media audiences increasingly visibly and powerfully shape organizational legitimacy.

Study 3

The third study (Stahel 2018) links micro to macro by conceptualizing aggressive acts as a tool used by members of social groups to silence journalists and thereby to protect the social identity and power of their own groups. The study is motivated by the assumption of amplified visibility and accessibility of journalists in the social media age. Potentially increasing online aggression against journalists might bias public media discourse in the long run, as it might silence

particular journalists. The study asks why some journalists seem to be more frequently attacked by their audiences than others. The macro level is acknowledged by assuming social groups with common norms and beliefs that strive for positive identities and power. On the micro level, group members may feel their social identity within a group to be threatened under particular conditions, such as if media coverage attacks their group. Once they perceive the image of their group as threatened by negative coverage, they may feel their social identity to be threatened. This motivates them to attack journalists so as to influence their behavior and ultimately protect the social identity and power of their group on the macro level. For group members, threats coming from journalists might include both macro-level properties such as journalists' ethnic groups and the power of their professional position and field of publishing and micro-level properties such as journalistic writing style. These properties connect to five central conditions that a politicized social identity approach (Tajfel and Turner 1979, Simon and Klandermans 2001) suggests increase threats to groups' social identity and power: evaluative settings, the salience of politicized identities, disruptions of strong place identity, evaluators that are powerful, and evaluators that belong to outgroups too similar to the ingroups. The journalists that are more frequently attacked are evaluative, publish on political topics, have a local focus, are powerful, and belong to similar outgroups.

This multilevel explanation of which journalists are attacked and why not only enlightens a practical problem, but also connects it theoretically to the sociological idea of society affecting individuals through social identities (Mead 1934, Tajfel and Turner 1979, Stryker and Burke 2000, Simon and Klandermans 2001). It contributes to literature on online aggression as it relates to more recent social identity explanations of online aggression (Coe, Kenski and Rains 2014, Nauroth et al. 2015, Rains et al. 2017). In contrast to them, it does not focus on processes among users of online platforms but the unequal targeting of actors within a highly public profession. The study also contributes to sociological and social psychology literature (Lawler et al. 1993, Stets and Burke 2000) by illustrating their explanatory potential for technology-mediated, aggressive behavior. Specifically, the proposed mutual contingency of social groups and online aggression suggests that aggression against journalists is a problem that emerges social-structurally rather than simply through people randomly and angrily targeting journalists.

1.5.3. Methodological contributions

In addition, the studies contribute behavioral data that are larger, broader, comparable, and include control cases, in contrast to most research on recent online aggression so far. The need for this has been stated by diverse scholars (Hutchens et al. 2015, Jane 2015, Runions et al. 2017, Koban et al. 2018). The methodological approach used in this dissertation increases the

external validity, generalizability, and reproducibility of results. The following section illustrates how this contribution is realized in each study.

Study 1

The first study (Rost et al. 2016) involves a data set that is so far, to my best knowledge, exceptional in the online aggression literature due to its richness (for one exception, see Vargo and Hopp 2017). The data set of 532,197 comments on 1,612 online petitions is much larger than most data sets of aggressive online comments, although very large data sets have been compiled recently (see section 1.2.2). The main contribution is that the data is both broader and comparable. Comments are drawn from only one online petition platform but includes a variety of topics from diverse societal fields such as politics, economy, and technology that have been submitted by online users all over Germany. Despite its topical and social-geographical breadth, comments are comparable because they share the same platform architecture. Commentators are thus assumed to be representative of a digital civil society: online users who actively contribute to a wide range of social-political issues. Aggressive comments can be compared to nonaggressive comments, and the online protests they are embedded in show a large spectrum of aggressiveness, from non-aggressive protests to highly aggressive firestorms. The availability of control cases for both comment and protest thus allows comparative conclusions. This also avoids the problem of selection for outcome (Sullivan 2014).

Study 2

The second study (Stahel and Rost 2018) contributes a rich data set and the measurement of relatively challenging moral legitimacy concepts to the online firestorm literature. It is, to my best knowledge, the first study that did not select a firestorm based on its labelling as a high-profile case by the media. The case was selected by choosing the largest and most aggressive online protest from the data set of online petitions used in the first study (Rost et al. 2016). This avoids selecting for outcome in that it is independent of labelling as a firestorm by news media (Sullivan 2014). The full data set of 45,997 aggressive and non-aggressive comments from a real-world firestorm is larger than those used in most former firestorm explorations. It provides external validity and allows systematic conclusions; this contrasts with most former studies on firestorms, which are either merely theoretical or use non-behavioral data. Comments were submitted from all over Germany. This renders its social significance greater than that of data for firestorms that are locally restricted. Further, the lexicon-based, content-analysis approach used is not only automated but also incorporates manual work. This allows the measurement of the complex concept of moral legitimacy among the comments. Such an approach goes beyond

former studies, whose automated text analyses allow only the simpler outcome variable of aggressive expressions to be measured.

Study 3

The third study's (Stahel 2018) major methodological contribution is the multivariate regression analysis of how often a representative sample of a public profession was attacked. Using inferential statistics on 530 journalistic survey participants goes beyond present online aggression studies, which so far have been limited to qualitative and descriptive approaches to exploring public targets of online aggression such as bloggers and journalists. In addition, it extends the few former studies that systematically explored who is exposed to online aggression, which have been largely limited to minors and adolescents. The data set is thus larger, comparable, includes control cases, and an adult population, in contrast to former related studies. The multilevel analysis of a large and representative sample of journalists thus contributes more sophisticated and generalizing conclusions about which representatives within groups are particularly frequently attacked.

1.6. In a nutshell: What did we learn and where to go from here?

Beyond the specific contributions of each study to research into why people comment aggressively through digital technologies, the overall dissertation provides a comprehensive and innovative approach to understanding and explaining online aggression. The proposed multilevel approach emphasizes its embeddedness in society. This socially broader approach effectively tackles the nature of online aggression as observed in more recent years. It thus adds knowledge to this complex, historically transforming yet highly topical phenomenon (see section 1.2.1.). It ultimately allows more valid predictions of what makes individuals and social groups comment aggressively as well as who is targeted most and why. The data-wise and methodological approaches presented here ensure valid empirical evidence to inspire future research.

Even more advanced and informative insights may be expected by extending this dissertation project in two directions. A first suggestion is to more explicitly and consistently apply and extend the multilevel approach, which is largely implicit in this dissertation, to study online aggression. For example, social-structural variables such as online aggressors' socio-economic status, education, and geographical region could be more systematically considered. Further, deducing and empirically testing links from micro to macro could be a complex but worthwhile undertaking. It could answer whether and how individual online aggressions indeed impact macro social structures and actors, notably norms, organizations, and public media discourse.

A second suggestion is to link online behavioral data such as hate comments with survey data of aggressors to even better apply multilevel explanations. This empirical approach is so far extremely rare, although surveys are optimal in exploring macro-level structural and associated beliefs (Maurer and Schmid 2010) and online comments represent valid micro behavior. Such designs pose practical and ethical challenges such as issues of privacy. However, it seems worthwhile to engage with these challenges, as they may provide a wealth of interesting insights into the as yet relatively uncharted territory of micro–macro interactions in online aggression. Finally, they would contribute to the more general question of how social behavior can be regulated in a digital future.

1.7. Personal contributions to co-publications

In the third study (Stahel 2018), the sole author, Lea Stahel, takes full responsibility for the content. The first study (Rost et al. 2016) and the second (Stahel and Rost 2018) were developed in co-authorship. Both the first and second study benefited from constant and mutual exchange between the doctoral student and her co-authors as well as other colleagues. Nevertheless, this section provides a rough orientation about the separate contributions of each author.

In the first study (Rost et al. 2016), Lea Stahel is the second author and Katja Rost and Bruno S. Frey are first and third authors respectively. As reported in the published paper (Rost et al. 2016), all authors cooperated in conceiving and designing the study and writing the paper. Lea Stahel and Katja Rost additionally collected, edited, analyzed, and interpreted the data, with Katja Rost contributing most.

In the second study (Stahel and Rost 2018), Lea Stahel is the first author and Katja Rost the second author. Lea Stahel is responsible for the majority of the development of the research question, theory, data collection, and analysis. She wrote the majority of the theory, analysis, and interpretation and collected and coded all variables of interest and all the control variables, including data from other sources such as news media coverage. Katja Rost collected and prepared the initial data set. In addition, she provided important inputs for the theoretical direction and data analysis. Finally, Katja Rost contributed in writing the paper, including restructuring, editing, and overall improving. Revisions were carried out in cooperation, to which Lea Stahel contributed the larger part.

1.8. Acknowledgements

Many people have contributed to the success of this dissertation, whom I would like to thank at this point. I start by thanking most sincerely my doctoral supervisor and primary review of my

thesis, Prof. Dr. Katja Rost. She enabled my lateral entry to the fascinating discipline of sociology and provided reliable support throughout the dissertation time. In addition, she continuously motivated and supported me in any opportunities as they came, both with the dissertation itself and with other activities such as the dissemination of my expertise. I also thank Prof. Dr. Sophie Mützel very much for her willingness to support this dissertation project as second reviewer.

In addition, I thank Fritz Schadow, the proprietor of the social media website www.openpetition.de, for permission to using all data received in the first and second studies. I also thank Ann-Sophie Gnehm for her help in programming in the second study. Finally, for the third study, I thank the journalists who were willing to be partners for exploratory interviews and who pretested the online survey.

Going back to the beginning of it all, I would also like to thank Prof. Dr. Christopher Cohrs, the supervisor of my master thesis, who assisted me in taking my first steps into academic life by motivating and supporting me in publishing my master thesis (Stahel and Cohrs 2015).

Furthermore, I would like to thank my expert colleagues for continuous and inspiring exchange. Here, I particularly want to mention Constantin Schön and Sebastian Weingartner. Last but not least, I want to thank my family, who believed in me and supported me.

2. Digital social norm enforcement: Online firestorms in social media[1]

Katja Rost[2], Lea Stahel[3], Bruno S. Frey[4]

Abstract

Actors of public interest today have to fear the adverse impact that stems from social media platforms. Any controversial behavior may promptly trigger temporal, but potentially devastating storms of emotional and aggressive outrage, so called online firestorms. Popular targets of online firestorms are companies, politicians, celebrities, media, academics and many more. This article introduces social norm theory to understand online aggression in a social-political online setting, challenging the popular assumption that online anonymity is one of the principle factors that promotes aggression. We underpin this social norm view by analyzing a major social media platform concerned with public affairs over a period of three years entailing 532,197 comments on 1,612 online petitions. Results show that in the context of online firestorms, non-anonymous individuals are more aggressive compared to anonymous individuals. This effect is reinforced if selective incentives are present and if aggressors are intrinsically motivated.

[1] This study is published under the same title in PLoS one (2016), 11(6). DOI: https://doi.org/10.1371/journal.pone.0155923
[2] Prof. Dr. Katja Rost, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland, E-Mail: katja.rost@uzh.ch
[3] Lea Stahel, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland. E-Mail: lea.stahel@uzh.ch
[4] Prof. Dr. Dr. h.c. mult. Bruno S. Frey, Center for Research in Economics, Management and the Arts, Südstrasse 11, 8008 Zurich, Switzerland. E-Mail: bruno.frey@bsfrey.ch

2.1. Introduction

Collective online aggression directed towards actors of public interest is an increasing phenomenon. While various types of social media have been involved in such online firestorms (e.g. content communities such as YouTube), blogs and social networking sites such as Facebook are outstanding triggers c. In 2011, Christian Wulff, the former federal president of Germany, was accused of corruption – claims that afterwards were rejected as unfounded although they promptly led to his resignation. The Wulff-affair was massively amplified by the negative word-of-mouth dynamics in social media. In 2013, the company Amazon was accused of the ill treatment of temporary workers. The Amazon-affair led to floods of negative comments on Amazon's Facebook profile. Firestorms also shake academia: In 2011, the former minister of defense of Germany, Karl-Theodor zu Guttenberg, was accused of plagiarism. These accusations triggered widespread online debates and ultimately led to the denial of his PhD and to his resignation.

The examples illustrate how online aggression has emerged from the private niche of limited email bullying to a publicly visible and relevant phenomenon. Dependent on the focus of the underlying research, the phenomenon of aggressive, offensive and emotional commenting in social media has been labeled flaming, cyberbullying, online harassment, cyber aggression, electronic aggression, toxic online disinhibition, trolling or, if the aggression resembles crowd-based outrage, online firestorms (Alonzo and Aiken 2004, Suler 2004, Buckels et al. 2014, Mehari et al. 2014, Pfeffer et al. 2014). In online firestorms, large amounts of critique, insulting comments, and swearwords against a person, organization, or group may be formed by, and propagated via, thousands or millions of people within hours (Pfeffer et al. 2014). Social media enable these unleashed phenomena (Suler 2004, Mishna, Saini and Solomon 2009, Mehari et al. 2014). They allow attacking everywhere at anytime with the potential for an unlimited audience. They raise the likelihood for hostile misinterpretations due to limited discursive action and social media's absence of nonverbal cues. They reduce the risk for feedback reactions because users can "sneak off" after the aggressive act.

The phenomenon of online aggression is not well understood despite the great deal of attention on hostile behavior in social media in both the mainstream media and the empirical literature (Ybarra and Mitchell 2004, Ybarra and Mitchell 2007, Mason 2008, Slonje and Smith 2008, Smith et al. 2008, Vandebosch and Van Cleemput 2008, Wolak, Finkelhor, Mitchell et al. 2008, Vandebosch and Van Cleemput 2009, Nocentini, Calmaestra, Schultze-Krumbholz et al. 2010, Kokkinos, Antoniadou and Markos 2014, Mehari et al. 2014). Most contributions are

descriptive and are conducted largely in the absence of theories (Kokkinos et al. 2014, Mehari et al. 2014). If contributions refer to theories they are mainly guided by traditional bullying research theory, more precisely by the massive amout of existing research concerned with cyberbullying among adolescents. Within this view, online aggression is understood as an irrational and illegitimate behavior that is caused by underlying personality characteristics, such as a lack of empathy and social skills, narcissism, impulsivity, sensation seeking, emotional regulation problems or psychological symptoms such as loneliness, depression, and anxiety (Sontag, Clemans, Graber et al. 2011, Kokkinos et al. 2014). Traditional bullying research theory, however, misses the point that in online firestorms, aggression happens in public, and not in private, social networks.

It therefore seems questionable whether bullying research theory is transferable to online firestorms. For example, a strong and commonly shared assumption within bullying research theory is that anonymity, understood as the degree to which a communicator perceives the message source as unknown and unspecified, promotes aggression through decreased inhibitions (Suler 2004, Ybarra and Mitchell 2004, Li 2007, Moore, Nakano, Enomoto et al. 2012, Hollenbaugh and Everett 2013). For online firestorms it suggests that negative, and particularly aggressive, word-of-mouth propagation in social media will weaken if real-name policies are introduced. In this article we show that this assumption is not necessarily true because the reverse effect can be obtained: Individuals have a strong motivation for being non-anonymous when being aggressive in social media. We explain this behavior pattern by social norm theory. Social norm theory may be a more appropriate theory to understand communication behavior in social media and to draw conclusions, for example, that real-name policies will not weaken online firestorms.

The remainder of this paper is structured as follows: the next section introduces social norm theory to understand aggressive behavior in a social-political online setting, and develops hypotheses. The subsequent sections explain the dataset, the measurements and the method, and present the empirical findings. We conclude with a discussion of the findings, research limitations and suggestions for further research.

2.2. A social norm theory on online firestorms

Social norms are fundamental to human behavior (Elster 1989, Guth and Napel 2006). Former literature defines norms as statements “that something ought or ought not to be the case” (Opp 2002, page 132), as institutionalized role expectations (Parsons 1964), or as becoming apparent if behavior attracts punishments (Homans 1950). In general, norms are mental representations

of appropriate behavior in society and smaller groups and, consequently, guide the behavior of individuals. Norms that are characterized as social "must be shared by other people and partly sustained by their approval and disapproval" (Elster 1989, page 99). Social norms are created intentionally because they promote the provision of a public good that benefits a collective, for example less pollution in a neighborhood due to less burning of leaves (Diekmann and Preisendörfer 1992), less harm to health through cessation of smoking (Opp 2002), or more fairness through income differentials (Fehr and Schmidt 1999, Rost and Weibel 2013). The public good view does not automatically imply that social norms are always beneficial for all persons concerned. In fact, many social norms exclude certain groups from public goods because they promote the interest of one subgroup, i.e., they serve "functions of inclusion and exclusion" (Elster 1989, page 108). For example, peer-group norms aim to strengthen cohesion within the group by offering group privileges (Elster 1989, Gunther, Bolt, Borzekowski et al. 2006).

To be sustainable, social norms need to be enforced, otherwise Olson's (1965) zero contribution holds: "if all rational and self-interested individuals in a large group would gain as a group if they acted to achieve their common interest or objective, they will still not voluntarily act to achieve that common or group interest" (Olson 1965: 2). Social norms are enforced by simple sanctions which trigger feelings of guilt and shame in the case of internalized social norms. Consequently, the mere expectation of sanctions, in turn, supports the enforcement (Elster 1989). Enforcement also happens through actual bilateral and multilateral costly sanctions where those who cause negative externalities are confronted with punishments and normative demands (Posner and Rasmussen 1999, Opp 2002). Linked to Olson's (1965) zero contribution, norm enforcement itself is a second-order public good: self-interested and utility-maximizing individuals do not naturally contribute to norm enforcement and may prefer free riding (Posner and Rasmussen 1999, Opp 2002). Ostrom (2000) however stresses how, in practice, contextual variables and the engagement of certain types of individuals determine whether collective action and cooperation is enhanced or discouraged. Similarly, Ellickson (Ellickson July 1999) emphasizes how norms may emerge or shift dependent on cost-benefit conditions or group composition. Also the presence, salience, or strength of social ties can explain individual variation in social-political engagement (McAdam 1986, McAdam and Paulsen 1993). For example, diffuse networks of weak bridging ties encourage mobilization through communicative advantage (Granovetter 1973). Specifically, research shows that Olson's (1965) second-order public good dilemma can be overcome if (1) norm enforcement is cheap, i.e., it occurs in low cost situations (Diekmann and Preisendörfer 1992, Rauhut and Krumpal 2008),

(2) additional benefits are provided to the norm enforcers that disproportionately motivate them compared to non-enforcers, i.e., selective incentives are present (Olson 1965, Opp 2002) and/or (3) if some individuals are present that are intrinsically motivated to enforce norms, i.e., some amount of altruistic punishment occurs (Bendor and Swistak 2001, Fischbacher et al. 2001, Fehr and Gächter 2002). In the following we elaborate these three conditions for social media to explain the phenomenon of online firestorms.

2.3. Online firestorms within a social norm theory

Aggressive word-of-mouth propagation in social media is the response to (perceived) violating behaviors of public actors. Public actors include, for example, politicians who disregard political correctness norms, corporations that violate human rights, or academics who violate scientific norms by engaging in plagiarism. In this view, online firestorms enforce social norms by expressing public disapproval with the aim of securing public goods, for example, honesty of politicians, companies or academics. The stunning waves of aggression typical for online firestorms can be explained by the characteristic features of social media that ideally contribute to the solution of the second-order public good dilemma of norm enforcement. Digital norm enforcement in social media is cheap, and selective incentives and intrinsically motivated individuals are present.

In social media, sanctioning norm violations occurs in low-cost situations. The basic idea of the low-cost hypothesis is that attitudes or preferences are more likely to guide individual behaviors when norm enforcement behavior is relatively cheap (Diekmann and Preisendörfer 1992, Rauhut and Krumpal 2008, Best and Kroneberg 2012). Evidence in various research fields supports this basic tenet (for an overview see Best and Kroneberg (2012)). For example, the voting paradox (Olson 1965), i.e., the fact that citizens participate in elections even though they are aware of the marginal influence of their vote, is often explained by referring to the low-cost hypothesis (Opp 2001). In social media, a number of factors contribute to such low-cost situations. First, social media mobilize former free riders because online criticism is monetarily inexpensive, hardly time-consuming and can be performed anywhere and anytime, compared, for example, to elaborate street protests (Mehari et al. 2014, Pfeffer et al. 2014). One example is the limited message length in the social media platform Twitter, which obliges communication to be short and quick. It is less astonishing that Twitter has been involved in most of the recent cases of online firestorms (Pfeffer et al. 2014). Second, in social media, people who are geographically completely removed from each other can assault each other verbally without fear of bodily harm. Nonverbal cues such as facial expression and physical

size are lacking, thus reducing the empathy of the aggressor and the impact of authority of the victims typically expressed by dress, body langugage, and social setting (Kiesler et al. 1984, Suler 2004, Mehari et al. 2014). Third, social media give ordinary people the power to communicate (perceived) norm violations to a very large audience (Harrington and Bielby 1995, Dennis 2008). The internet re-creates village-like interconnectedness within a global, pluralistic society by crossing local, or even national, boundaries due to unrestrained information flow (Castells 2012). To compare, while aggressive norm enforcement is a rare behavior in the non-digital context (Brauer and Chekroun (Brauer and Chekroun 2005) found that max. 4% of bystanders aggressively sanction daily deviant behavior by insulting or aggressive shouting), we should observe it more frequently in the digital social media context for the reasons given above.

> *Hypothesis 1: Provided that a social-political issue finds its way into social media platforms, online aggression takes place more frequently than in the non-digital context because sanctioning of (perceived) norm violations occurs in low-cost situations.*

In social media, selective incentives that benefit a latent group of norm enforcers are disproportionally present (Olson 1965, Opp 2002). Individuals only bear the costs of norm enforcement if the potential benefits of their actions exceed the costs (Fehr and Fischbacher 2004). Selective incentives translate resentment for norm breaching into action in situations where it is unclear whether a necessary critical mass of other norm enforcers will join the action. In such situations, cost sharing cannot be expected, nor can clear benefits from norm enforcement, such as an actual behavioral change by the accused person or organization, be predicted. In the case of selective incentives, individuals participate in collective action in response to salient private benefits (Ginges and Atran 2009). Whether individuals are able to reap selective incentives is dependent on the issue at stake and on certain individual or group characteristics. Social media contribute to the presence of selective incentives by enhancing the salience of private benefits. In social media, for example, highly controversial topics are debated. Social media are, in addition, highly influenced by the multiplication of cross-media dynamics, for example by public scandals taken up or created by news media leading to comments in social media. Broad public discussions and connections to public scandals give credible signals that a norm infringement at the expense of a latent interest group – be it the group an individual belongs to or identifies with – has occurred (Myers 2000).

Hypothesis 2: Online aggression in social media is encouraged by salient selective incentives, for example, in highly controversial topics or in topics connected with public scandals.

Social media ensure that a high amount of intrinsically motivated actors are present. Individuals engage in costly norm enforcement if they have an intrinsic desire to "make the world a better place" (Lee and Tedeschi 1996, Salmivalli, Lagerspetz, Bjorkqvist et al. 1996, van Stekelenburg, Klandermans and van Dijk 2011). This type of norm enforcement has been intensively discussed as "altruistic punishment", i.e., individuals punish, although the punishment is costly for them and yields no material gain (Bendor and Swistak 2001). Altruistic punishment is driven by strong negative emotions towards the norm defector (Fischbacher et al. 2001, Henrich, Boyd, Bowles et al. 2001, Fehr and Gächter 2002) and by people's perception of a state of affairs as illegitimate (Tajfel and Turner 1979, Feather and Newton 1982, Weiss, Suckow and Cropanzano 1999, Klandermans 2003, Van Zomeren, Spears, Fischer et al. 2004). Strong intrinsic motivation, however, is only likely to encourage participation if it is reinforced by organizational or individual ties (McAdam and Paulsen 1993). This requirement is given in the infrastructural setting surrounding online firestorms. The technical mechanisms of social media such as newsletters, newsgroups, followers, or social media sharing ensure that intrinsically motivated individuals are optimally informed about cases that, in their view, represent offenses against existing social norms. Beyond this, they provide opportunities to tackle these norm violations by commenting on them.

Hypothesis 3: Intrinsically motivated actors encourage online aggression in social media.

2.4. The non-anonymity of negative word-of-mouth dynamics in social media

In social media, people can hide or alter their identity. They may either comment by providing no name or at least not their real name, i.e., a (random or stable) pseudonym. Existing literature on online behavior hypothesizes that such online anonymity is one of the principle factors that decreases inhibitions, increases self-disclosures and therefore promotes online aggression (Suler 2004, Ybarra and Mitchell 2004, Li 2007, Moore et al. 2012, Hollenbaugh and Everett 2013). This causal mechanism is also assumed by social media consultants who attempt to explain online firestorms (Bishop 2014).

In general, anonymity produces the "stranger on a train" phenomenon, wherein people share intimate self-disclosures with strangers as they do not expect a reunion and hence do not fear any risks and constraints (Bargh, McKenna and Fitzsimons 2002). To that effect, "when people

have the opportunity to separate their actions online from their in-person lifestyle and identity, they feel less vulnerable about self-disclosing and acting out" (Suler 2004, page 322). With regard to heightened aggression and inappropriate behavior, psychosocial motives exist for being anonymous (Moore et al. 2012). Anonymity first detaches from normative and social behavioral constraints (Patchin and Hinduja 2006). Second, it allows to bypass moral responsibility for deviant actions (Suler 2004). Third, it reduces the probability of social punishments through law and other authorities (Li 2007). Fourth, it triggers an imbalance of power which limits the ability of the victim to apply ordinary techniques for punishing aggressive behavior (David-Ferdon and Hertz 2007). Fifth, it gives people the courage to ignore social desirability issues (Suler 2004) and finally, it encourages the presentation of minority viewpoints or viewpoints subjectively perceived as such (DeSanctis and Gallupe 1987, Dennis 1996, Gopal and Prasad 2000, McLeod 2000, Dennis and Garfield 2003).

Former research has concluded that the possibility for anonymity in the internet fosters aggressive comments. It is assumed that online aggression is driven by lower-order moral ideals and principles and, consequently, people feel ashamed to aggress under their real names. However, the empirical evidence for such a link is scarce and no definitive cause-effect relationship has evolved. Studies suggest that anonymity only increases online aggression in competitive situations (Hughes and Louw 2013), that anonymity does not increase online aggression but does increase critical comments (Reinig and Mejias 2004), or that the effect of forced non-anonymity on the amount of online aggression is a function of certain characteristics of user groups, e.g. their general frequency of commenting behavior (Cho and Kim 2012).

The former conceptualization of online aggression is rather narrow, in particular for aggression in social media. According to social norm theory, in social media, individuals mostly use aggressive word-of-mouth propagation to criticize the behavior of public actors. As people enforce social norms and promote public goods, it is most likely that they perceive the behavior of the accused public actors as driven by lower-order moral ideals and principles while that they perceive their own behavior as driven by higher-order moral ideals and principles. From this point of view there is no need to hide their identity.

Furthermore, aggressive word-of-mouth propagation in a social-political online setting is much more effective if criticism is brought forward non-anonymously. This is due to the fact that non-anonymity inceases the trustworthiness of the masses of weak social ties to which we are linked, but not necessarily familiar with, in our digital social networks. Trustworthiness of former firestorm commenters encourage us to contribute ourselves. First, non-anonymity is

more effective as the credibility of sanctions increases if individuals use their real name (Dennis 1996, Haines, Hough, Cao et al. 2014). Anonymity makes "information more suspect because it [is] difficult to verify the source's credibility" (Dennis 1996, page 450). This removes accountability cues and lets one assume that individuals present socially undesirable arguments (Prentice-Dunn and Rogers 1982, Haines et al. 2014). Second, the views of non-anonymous individuals are given more weight: "Just as people are unattached to their own statements when they communicate anonymously, they are analogously unaffected by the anonymous statements of others" (McLeod 2000, page 197). Anonymous comments have less impact on the formation of personal opinions (McLeod 2000, Sassenberg and Postmes 2002), on the formation of group opinions (Haines et al. 2014), and on final decision making (Stanley and Weare 2004). Third, anonymity lowers the identification with, support of, and recognition by, kindred spirit (Douglas and McGarty 2001). In anonymous settings, individuals cannot determine who made a particular argument, how many different people expressed similar arguments, whether a series of arguments are all coming from the same person, or the degree to which other commenting individuals are similar to oneself (Gutwin and Greenberg 2002, Hayne, Pollard and Rice 2003, Lee 2007, Haines et al. 2014). Anonymity filters out cues that communicate social identity, cues that are necessary to characterize comments by others (Cooper and Haines 2008, Haines et al. 2014), to identify with individuals in social comparison processes (Haines et al. 2014) and to coordinate group interactions (Gutwin and Greenberg 2002). Finally, anonymity reduces the benefit to be positively evaluated by others (Valacich, Jessup, Dennis et al. 1992, Pinsonneault and Heppel 1998). Studies show that exclusively anonymous conditions induce little mobilization because anonymity excludes the benefit of recognition by others (Andreoni and Petrie 2004).

From a social norm point of view, the arguments suggest that aggressive word-of-mouth propagation in a social-political online setting takes place non-anonymously. People have a strong feeling to stand up for higher-order moral ideals and principles. Commenting anonymously is a costly, wasteful behavior, as sanctions are less credible, create less awareness, less support and offer few benefits. These considerations make particular sense in the usual setting of firestorms, namely social media where usually, weak social ties are clustered around ideologically like-minded networks. Such networks likely support non-anonymous aggressive sanctions that confirm their worldview.

> *Hypothesis 4: In a social-political online setting, non-anonymous individuals, compared to anonymous individuals, show more online aggression.*

As stated earlier, norm enforcement is fostered if selective incentives and intrinsically motivated actors are present. Consequently, if social norm theory is an appropriate theory for online aggression in a social-political online setting, these groups in particular should give more weight to the benefits of non-anonymous aggressive word-of-mouth propagation. Simultaneously, they give less weight to potential risky consequences such as being subject to deletion, banned from websites, formally convicted by the accused actor for defamation of character and/or damage to reputation, or informally sanctioned by social disapproval from online or offline individuals (Tichy 2013).

> *Hypothesis 5: In a social-political online setting, in situations that offer selective incentives, compared to situations without selective incentives, more online aggression by non-anonymous individuals is observed.*
>
> *Hypothesis 6: In a social-political online setting, intrinsically motivated aggressors (i.e. aggressive commenters), compared to aggressors without intrinsic motivation, show more online non-anonymous aggression.*

2.5. Materials and Methods

2.5.1. Sample

We test the hypotheses with a census of a major social media platform concerned with public affairs. We analyze all comments on online petitions published at the German social media platform www.openpetition.de between May 2010, the launching of the online portal, and July 2013 (for permission of using the data, see document 1 in the Appendix). Online petitions exemplarily include protests against pay-scale reform of the German society for musical performing and mechanical reproduction rights called GEMA (305,118 signers), against the enforcement to finance public service media (136,010 signers), against the closing of the medical faculty at the University Halle (58,577), or for the resignation of an Austrian politician (9,196 signers) or the Bavarian minister of justice (6,810 signers). Online petition platforms seem very suitable to investigate the phenomenon of negative word-of-mouth in a social-political online media setting. First, online petitions are concerned with public actors and public affairs, for example, internet security, misbehavior of firms, politicians, or academics, public spending, tax issues, animal protection, etc., and thus provide a central location where public norms are negotiated. Second, online petition platforms are prototypical social media platforms: everybody is allowed to participate and create content for any kind of topic, and the debates and comments are publicly visible. Third, qualitative evidence suggests that many popular firestorms have been triggered or have been surrounded by online petition platforms, for

example the Deutsche Telekom firestorm in 2013, or the firestorm leading to the displacement of the German Federal President Christian Wulff in 2011. Fourth, online petition platforms are concerned with real-life cases. Many former studies are based on artificial laboratory experiments to study negative word-of-mouth behavior on the internet. Finally, online petition platforms cover a wide range of public issues and affairs, implying lower selection biases as compared to case studies about online firestorms (such as in Pfeffer et al. (2014)).

The final dataset includes 532,197 comments on 1,612 online petitions. There were a total of 3,858,131 signatures over the 1,612 petitions between 2010 and 2013, with detailed information about the wording of the comment, the commenters, the signers and the petition. The dataset was provided to the authors in an anonymous form by the platform owner. For each signer and commenter, however, the dataset indicated whether he/she had originally contributed anonymously (=1) or non-anonymously (=0). For this study, no approval of any ethics committee was sought because all data are publicly accessible on www.openpetition.de and no names of signers or commenters can be tracked and identified in the dataset. In order to prepare the dataset in accordance with our theory, we rely on a mixed-method big-data approach. For many variables we use a qualitative approach to arrive at meaningful quantitative measurements.

The present dataset allows us to exclude two biases which, in other studies, frequently affect findings on relations between anonymity and aggression. First, there was no active intervention in the ratio of anonymous and non-anonymous aggressive comments in the dataset. In the period of data collection, the platform owner did not moderate the comments on his own initiative. However, he reacted by deleting selected inappropriate comments when the user community reported them. According to the platform owner, a deletion was independent of whether the inappropriate comment was provided anonymously or not, as he explicitly considered this difference as irrelevant to liability issues. Second, we may also exclude any bias stemming from differing legal jurisdictions: Potential legal implications for identified aggressors are the same across the entire study. In Germany, the jurisdiction on defamation and insult is part of the federal law (Bundesministerium der Justiz und für Verbraucherschutz and juris GmbH 2016), i.e., as the entire study pertains to the same legal jurisdiction, all defamatory or aggressive commenters across all German states face the same potential costs for their actions.

2.5.2. Measurement of Variables

We measure online aggression in the following manner. In general, inconsistency in the operationalization of online aggression dominates research (Joinson 2007). Operationalization

includes impolite statements, swearing, flirting, exclamations, expressions of personal feelings, use of superlatives (Kiesler, Zubrow, Moses et al. 1985) to profanity, typographic energy (e.g. exclamation marks), name calling, swearing, and general negative effect (Reinig and Mejias 2004, Joinson 2007). We rely on the definition of online aggression in firestorms, i.e., large amounts of critique, insulting comments, and swearwords against a person, organization, or group formed by, and propagated via, social media platforms (Pfeffer et al. 2014). Accordingly, we measure online aggression by direct offenses within the comments on online petitions (e.g. "I hate GEMA, complete morons and exploiters", ID469090), swearwords (e.g. "Fuck that Shit!", ID477368), and expressions of disgust or contempt (e.g. "The deportation policies of German authorities is commonly a disgusting, repulsive and inhuman mess!", ID418089). Expressions of disgust and contempt are typical responses to morally offensive behavior (Hutcherson and Gross 2011). Importantly, even from the outside perspective, we confidently evaluate these expressions to be intended as aggression. This is because we do not expect close relationships or shared, subcultural interactional norms between the commentator and the targeted actor in petitions, in contrast to profane language between friends representing covert closeness and not aggression (O'Sullivan and Flanagin 2003).

To systematically collect online aggression, we compile a list of frequently used swearwords from synonym reference books and online databases of swearword collections (e.g. https://www.schimpfwoerter.de). This approach corresponds to previous studies that count aggressive postings by using a pre-defined set of aggressive words (such as in Cho and Kim (2012)). Then, we disaggregate the 532,197 comments into single words and count them. Frequently occurring words are manually checked and classified as online aggression if applicable. Subsequently, we exclude all words that can be used for different meanings, for example, as swearwords or as terms for animals. These steps led to a final list of 1,481 words that express offenses, swearwords, and disgust. Using this final list of aggressive expressions, we count the amount of online aggression in each comment. Subsequently we qualitatively check the appropriateness of our approach by comparing subsamples of comments with our quantitative measurement. We take the logarithm added by 1 to create an approximate normal distribution of the variable.

2.5.3. Independent variables

Anonymity is measured in the following way: Before online users sign a petition and subsequently formulate a voluntary comment, they are requested to provide their real names and addresses. In regard to public visibility, they are given the choice to allow their real name to be published or to remain anonymous, i.e., only the postal code is visible to other users (0 =

non-anonymous, 1 = anonymous). Although the theoretical possibility of using pseudonyms does exist, we expect that commenters' incentive for pseudonyms is low. This is because anonymity complies with the hidden name option and petition organizers may classify the signature of pseudonyms as invalid.

Controversy that accompanies a petition is measured by the level of debate. Each petition provides the opportunity to start a debate on the petition homepage, a tool used in most petitions by supporters and opponents. A debate is structured by denoted pro- and contra-arguments, i.e., by arguments that underpin or oppose the petition's concerns. Only arguments that differ in their content from formerly mentioned arguments are additionally incorporated. Within the pro- and contra-sections, commenters are allowed to oppose arguments by adding sub-replies (pro-reply-/contra-reply-arguments). More controversial topics lead to a higher diversity of pro-, contra-, pro-reply- and contra-reply-arguments. Thus, to measure controversy, we construct a Herfindahl index by taking the percentage of arguments within each category, i.e., pro-/contra-/pro-reply-/contra-reply-arguments, squaring it, adding them together and subtracting the final result from 1. The index measures the controversy that surrounds the topics of petitions from no controversy (= 0) to a maximum of controversy (= 1).

To identify scandals, we measure whether the accusation against an actor forwarded by a petition, for example corruption of a politician, is covered and framed as scandal by traditional news media (1 = yes / 0 = no). We define keywords that describe the content and concerns of the petition. In the database LexisNexis we search for whether these keywords are associated with the term "scandal" in the German-speaking media within a time period of one year before the starting date of each petition.

To measure actors' intrinsic motivation, we operationalize fairness perceptions of commenters. We compile a list of 579 expressions frequently used in ideological discourses that indicate fairness issues, for example, expressions such as "injustice" or "unfair". In addition, we use synonym reference books and databases, manually check frequently occurring words within comments and exclude ambiguous words. For each commentator we count the amount of intrinsic motivation by taking the sum of fairness words in the comment. We take the logarithm, added by 1, to create an approximate normal distribution of the variable.

2.5.4. Control variables

We control for factors that influence the amount of online aggression.

The length of comment is measured by the total number of words in a comment. Longer comments are more likely to entail more aggression.

The time period between opening a petition and submitting a comment is included because the time point of comment submission may influence commenters' level of aggression. Aggression may either take place in the very beginning, because most signatures and comment activity in petitions are submitted within the first days (Hale, Margetts and Yasseri 2013), or alternatively, in advanced stages, in the case where a petition experiences a boost due to revived public debate. We measure how many minutes after petition opens that a comment has been submitted.

The number of protesters having signed is included because larger protests are likely to attract more online aggression. We measure how many individuals sign a particular petition and consequently match this data with the comments on a certain day. The median of protesters amounts to 76 signers per day with a maximum of 2,926 signers per day. We take the logarithm of the number of protesters to create an approximate normal distribution of the variable.

The status of the accused may also influence online aggression. Theoretically, public actors with a high social status may be either protected from sanctions as they have more resources to reply to punishments by even more painful punishments, or, to the contrary, they can attract sanctions because they are also more vulnerable than lower status actors (Wahrman 2010). In practice, high status celebrities or politicians may also refrain from suing laypersons as it is counterproductive to their reputation. To take these complex influences into account, we control for the status of the accused. As a proxy for social status of the accused public actors, we collect the number of Google hits for the accused's name (1 = <1000; 2 = <10,000; 3 = <100,000; 4 = <500,000; 5 = <1,000,000; 6 = >1,000,000). Google hits tend to reflect social status. To decrease measurement errors, for example due to actors sharing the same name, we additionally check whether the accused is listed in the German online encyclopedia Wikipedia (0 = no entry, 1 = entry in article's subtitle, 2 = entry as main article). Wikipedia exclusively lists actors with a minimum public status. We add both variables and take the logarithm of the mean value.

We measure also whether the accused is a natural person or a legal entity. Legal entities professionally monitor the internet for defamation and gather more resources to fight accusations than do natural persons. To avoid that commenters anticipate differing costs for their aggressive behavior dependent on whom the accused actor is, we control for this factor. Two independent coders manually check whether the target is a natural person such as a scientist or politician (= 1) or a legal entity such as a government or an organization (= 0). In 4% of the petitions, the target is a natural person and not a legal entity.

The anonymity of the social environment of commenters measures the anonymity of the environment in which commenters live. This may influence how much aggression is expressed

(Cammaerts and van Audenhove 2005). Less anonymous villages with tight social control likely increase sanctioning costs. As a proxy for the anonymity of commenters' social environment, we measure the size, i.e., the number of inhabitants, of the city or village in which commenters live. The postal codes of each signer are aggregated such that individuals living in the same city or village are merged. The dataset includes 23,977 cities and villages. We count the number of signers for each city or village, and by random checking, we find that the correlation of the number of signers within a postcode region, and the de facto size of this region, is 0.92, validating our proxy. We allocate the size of residence variable to all signers and commenters. Bigger values indicate that commenters originate from more anonymous environments.

The regional scope of a protest is measured because issues of broad public relevance may attract more aggression. We measure the regional diversity of a petition by constructing a Herfindahl index ranging from no regional diversity (= 0) to a maximum of regional diversity (= 1). Signers are assigned to different German federal states on the basis of residential postal codes. We take the percentage of signers within each federal state, square it, add them together, and subtract the final result from 1.

The success of a petition is measured because successful petitions potentially deal with more relevant topics, which may indirectly influence the amount of online aggression. A petition is considered successful if the petition initiator defines the petition goals to be achieved in full or at least in part (1 = yes; 0 = no).

The petition motive may influence the amount of online aggression. Using a petition's title and leading text, two independent coders classify the petitions with regard to their underlying motives by using the classification by Reiss (2004). Five major concerns are identified, namely idealism/fairness (42%), income/costs (19%), security/social order (13%), autonomy/self-determination (14%), and quality of life/competences (52%). Multiple assignments of petitions are possible. Idealism/ fairness serves as the reference group in the regression models.

Similarly, the petition topic may influence anonymity considerations and the amount of aggression. Depending on the societal area, be it the economy, politics, or culture, accused actors may differ in their thresholds of wanting to sue aggressive online commenters. Commenters may anticipate these thresholds and the related differing costs of being aggressive. This in turn affects commenters' actual behavior. Using a petition's title and leading text, two independent coders classify the petitions with regard to their underlying topics using the functional systems of a society (Oehmer 2011). Six major topics are identified, namely society

(41%), arts (20%), economics (13%), politics (8%), media (8%), and environment and animal protection (8%). Multiple assignments of petitions are avoided. Society, including topics such as sport or solidarity, is the most general category and serves as reference group in the regression models.

For the summary of the descriptive statistics and bivariate correlations of the former variables, see Table 5 in the Appendix.

2.5.5. Methods

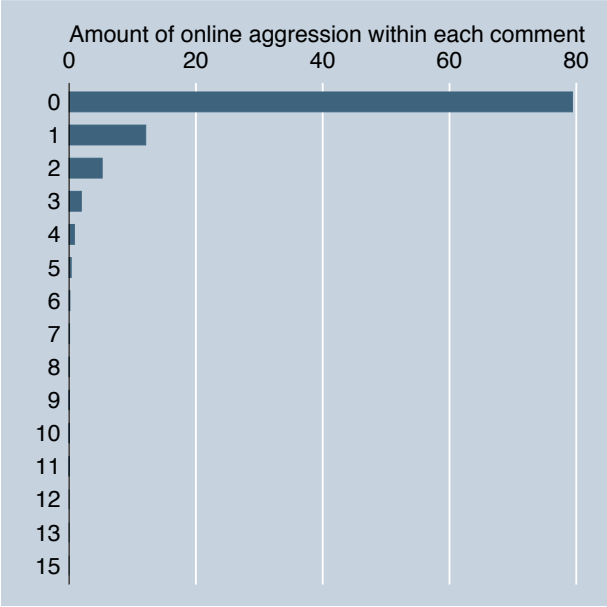
We apply random-effects and fixed-effects models to predict online aggression in petitions.[5] In both models the comments are grouped on the petition level. The random-effects model keeps within- and between-petition variation in the analysis. We assume that petitions vary not only within, but also between, each other, for example because some petitions have many supporters while other petitions have only a few supporters, or because of differences in the underlying goals and motives. We analyze whether online aggression within and between petitions changes when other variables within and between the petitions change. The fixed-effects model keeps only within-petition variation in the analysis. We also analyze whether the aggression within petitions changes when other variables change, for example the anonymity of commenters, the amount of intrinsic motivation or the amount of selective incentives within the petitions. Many variables of our dataset are time-invariant, i.e., constant petition features that do not vary on the petition level. In the fixed-effects model these variables are omitted. Both models have advantages as well as disadvantages. The fixed-effects model excludes all random noise between the petitions and is therefore often preferred as the golden standard. However, differences between the petitions, for example the number of supporters, may also be important in explaining negative word-of-mouth behavior within petitions. This speaks in favor of the random-effects model. We therefore apply both models and compare the results. We additionally run alternative conceivable models for the data structure, for example, logistic regression, Poisson regression, or negative binomial regression for panel data, as our dependent variable is (if not transformed) a count variable, or can be transformed into a binary variable that indicates whether a person is an aggressor or not. The results are similar with the results that follow and will therefore not be presented here.

[5] All data and syntax is available from the Inter-university Consortium for Political and Social Research:
Rost, Katja; Stahel, Lea; Frey, Bruno S. Online Petition Data: 2010-2013 (Germany). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-05-11. http://doi.org/10.3886/E72764V1

2.6. Results

In accordance with Hypothesis 1, the data substantiate that online aggression in social media is a more frequent phenomenon than in the non-digital context. In the analyzed online petition platform, we find 197,410 aggressions according to our definition. 20.62% of all comments entail a minimum of one aggressive expression (Figure 2.1). In 9% of all comments we find two, up to fifteen, aggressive expressions. On the petition level, only 11% of all petitions include no aggressions. 34% include a negligible amount of aggressions from 1, up to 10. 37% include 11 up to 100 aggressions. 16% include 101 up to 1,000 aggressions. 2% include 1,001, up to 25,360, aggressions. Even if the prevailing majority of commenters make no use of aggressive language in social media, the numbers demonstrate that online aggression occurs not only in a vanishing minority of comments or petitions (compared to the observed vanishing minority of max 4% of bystanders aggressively sanctioning in the non-digital context (Brauer and Chekroun 2005). This supports the claim that in social media, aggressive sanctioning behavior is a relatively frequent phenomenon because it takes place in low-cost situations.

Figure 2. 1. Observed amount of online aggression per comment



We now move to the presence of selective incentives and intrinsically motivated actors in social media. The descriptive findings show that 47% of all petitions are accompanied by a highly controversial debate, 6% of the petitions are associated with a scandal in news media, and 26% of the commenters are motivated by fairness concerns. Social media thus indeed seem to offer an environment in which the second-order public good dilemma of norm enforcement can be overcome. Whether these conditions indeed contribute to norm enforcement is tested in Tables 1 and 2.

Table 1. Predicted amount of online aggression dependent on the anonymity of aggressors (random-effects regression)

	Model 1				Model 2			
Y: Amount of online aggression (log)	Coef.	Std.Err.	z	P>|z|	Coef.	Std.Err.	z	P>|z|
Anonymity	-.02	.00	-13.10	***	.00	.00	-.35	
Controversy of accusation	.04	.01	4.45	***	.05	.01	4.86	***
Accusation is connected to a scandal	.02	.01	2.16	*	.03	.01	2.38	*
Intrinsic motivation (log)	.01	.00	12.17	***	.02	.00	12.15	***
Anonymity x Controversy					-.02	.01	-3.01	**
Anonymity x Scandal					-.01	.00	-3.00	**
Anonymity x Intrinsic motivation					-.01	.00	-3.19	**
Length of comment in words	.00	.00	114.09	***	.00	.00	114.13	***
Time of comment after petition opening	.00	.00	-3.31	**	.00	.00	-3.30	**
Number of protest participants (log)	.00	.00	-.35		.00	.00	-.33	
Scope of protest	.03	.01	3.38	**	.03	.01	3.39	***
Success of the petition	.01	.01	.71		.01	.01	.70	
Status of the accused (log)	.00	.01	-.38		.00	.01	-.43	
Accused is a natural person (vs. legal entity)	.05	.01	4.03	***	.05	.01	4.03	***
Anonymity of social environment of aggressors (log)	.00	.00	-5.69	***	.00	.00	-5.68	***
Motives: Income/minimization of costs	-.01	.01	-1.28		-.01	.01	-1.30	
Motive: Security/social order/traditional values	.01	.01	1.29		.01	.01	1.29	
Motive: Independence/self-determination	.00	.01	.05		.00	.01	.05	
Motive: Increasing life quality and competence	-.06	.01	-8.65	***	-.06	.01	-8.69	***
Topic: Art/culture/education	-.01	.01	-1.25		-.01	.01	-1.26	
Topic: Economics	.02	.01	1.97	*	.02	.01	1.98	*
Topic: Politics	.00	.01	.13		.00	.01	.15	
Topic: Media	.05	.01	4.01	***	.05	.01	4.01	***
Topic: Environmental and animal welfare	.05	.01	4.37	***	.05	.01	4.40	***
Constant	.06	.02	3.88	***	.06	.02	3.70	***
Number of observations			532196				532196	
Number of groups			1568				1568	
R-square (between)			12.69%				12.70%	
Wald chi2			15031.07	***			15066.10	***

Legend: †< p .1, *< p .05, **< p .01, ***< p .001

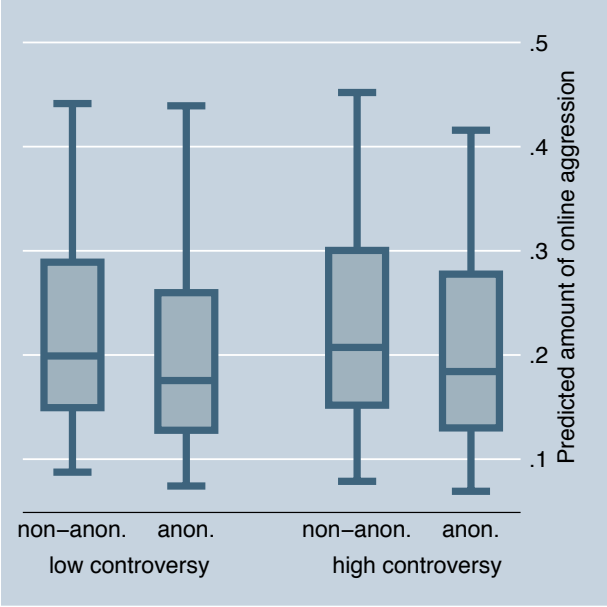
Table 2. Predicted amount of online aggression dependent on the anonymity of aggressors (fixed-effects regression)

	Model 1				Model 2			
Y: Amount of online aggression (log)	Coef.	Std.Err.	z	P>|z|	Coef.	Std.Err.	z	P>|z|
Anonymity	-.02	.00	-13.14	***	.00	.00	-.29	
Controversy of accusation	(drop.)				(drop.)			
Accusation is connected to a scandal	(drop.)				(drop.)			
Intrinsic motivation (log)	.01	.00	11.79	***	.02	.00	11.82	***
Anonymity x Controversy					-.02	.01	-3.07	**
Anonymity x Scandal					-.01	.00	-3.00	**
Anonymity x Intrinsic motivation					-.01	.00	-3.18	**
Length of comment in words	.00	.00	114.00	***	.00	.00	114.04	***
Time of comment after petition opening	.00	.00	-3.63	***	.00	.00	-3.64	***
Number of protest participants (log)	.00	.00	-.31		.00	.00	-.29	
Status of the accused (log)	(drop.)				(drop.)			
Scope of protest	(drop.)				(drop.)			
Success of the petition	(drop.)				(drop.)			
Accused is a natural person (vs. legal entity)	(drop.)				(drop.)			
Anonymity of social environment of aggressors (log)	.00	.00	-5.79	***	.00	.00	-5.77	***
Motives: Income/minimization of costs	(drop.)				(drop.)			
Motive: Security/social order/traditional values	(drop.)				(drop.)			
Motive: Independence/self-determination	(drop.)				(drop.)			
Motive: Increasing life quality and competence	(drop.)				(drop.)			
Topic: Art/culture/education	(drop.)				(drop.)			
Topic: Economics	(drop.)				(drop.)			
Topic: Politics	(drop.)				(drop.)			
Topic: Media	(drop.)				(drop.)			
Topic: Environmental and animal welfare	(drop.)				(drop.)			
Constant	.11	.00	33.16	***	.11	.00	32.90	***
Number of observations			532196				532196	
Number of groups			1568				1568	
R-square (within)			2.70%				2.70%	
F-value			2449.47	***			1636.62	***

Legend: †< p .1, *< p .05, **< p .01, ***< p .001

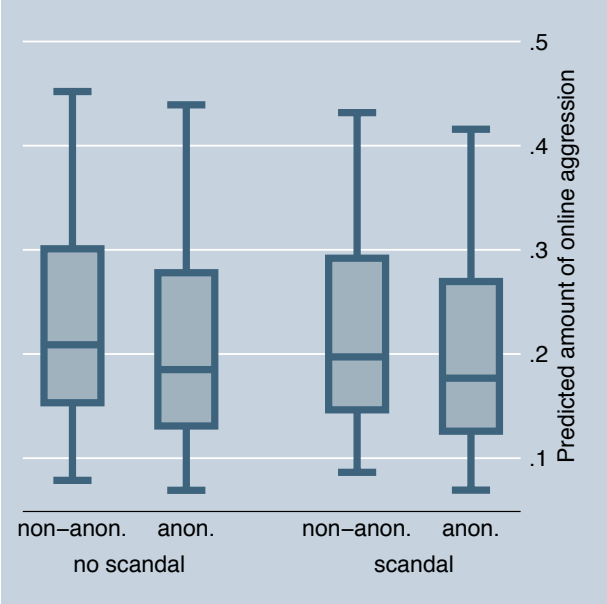
The random-effects model in Table 1, Model 1, confirms that situations that offer selective incentives, i.e., a petition is accompanied by a highly controversial debate or is connected with a scandal in news media, significantly encourage online aggression in comments. This preliminarily supports Hypothesis 2 (for the size of the effects see Figures 2.2 and 2.3). The fixed-effect model in Table 2 entails no results for selective incentives because petition-invariant effects are dropped. Further, the random-effects as well as the fixed-effects models in Tables 1 and 2, Model 1, preliminarily support Hypothesis 3: online aggression is encouraged by intrinsically motivated actors as compared to individuals without fairness concerns (for the size of the effects see Figures 2.4 and 2.5).

Figure 2. 2. Online aggression dependent on controversy and anonymity (random-effects)



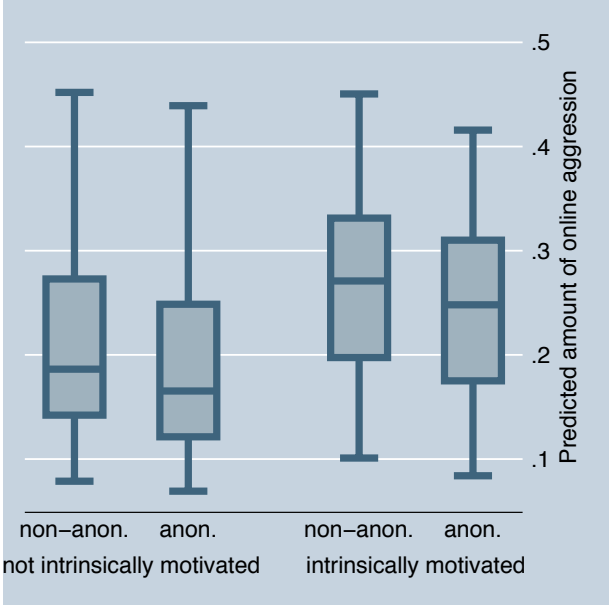
Predictions of Table 1, Model 2.

Figure 2. 3. Online aggression dependent on scandal and anonymity (random-effects)



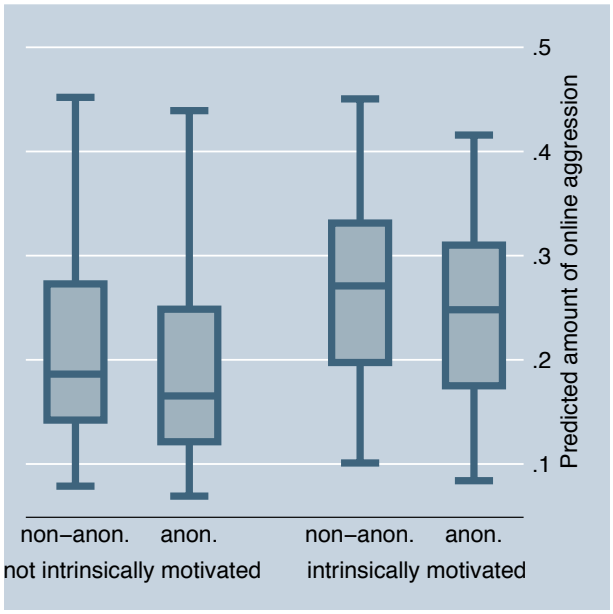
Predictions of Table 1, Model 2.

Figure 2. 4. Online aggression dependent on intrinsic motivation and anonymity (random-effects)



Predictions of Table 1, Model 2.

Figure 2. 5. Online aggression dependent on intrinsic motivation and anonymity (fixed-effects)



Predictions of Table 2, Model 2.

Building on the view that social media today are a major channel for digital social norm enforcement, which until now is not rejected by the data, Hypothesis 4 assumes that online aggression takes place non-anonymously. Aggressive commenters have nothing to hide: they stand up for higher-order moral ideals and principles. The goal of norm enforcement can be reached most effectively if sanctions are forwarded non-anonymously because they are credible, create awareness, support, and offer benefits. The descriptive statistics show that only 29.2 % of all commenters prefer to remain anonymous. Anonymity of commenters is thus a characteristic feature of social media; however, a vast majority still comments under their real names. The results in Tables 1 and 2, Model 1, show the impact of commenters' anonymity to predict online aggression in comments. In line with Hypothesis 4, both the random-effects and fixed-effects models show that more online aggression is obtained by non-anonymous commenters and not by anonymous commenters.

Exemplarily, we present three of the most aggressive comments by non-anonymous commenters: "Silly, fake, inhuman and degrading, racist, defamatory and ugly theses like those of Sarrazin (author's note: a former German politician) have no place in this world, let alone in the SPD (author's note: Social democratic party). Sarrazin certainly has no business in the Social democratic party and should try his luck with the Nazis" (ID352216); "HC Strache (author's note: Austrian politician) has an evil, inhuman character, he lies and tries to persuade other people of wrong ideas." (ID284846); "These authorities are mostly no people, but §§§- and

regulatory machines! I detest authorities – with my 67 years' life experience after all!" (ID418089).

Figures 2.6 and 2.7 illustrate the size of the effect as predicted in the random- and fixed-effects regressions. The average effect of anonymity on aggression becomes sharper in the fixed-effects model. The random-effects model additionally illustrates that many of the very aggressive commenters appear non-anonymously. Overall, the effect size is small. However, the data clearly show that social norm enforcement, and not as popularly assumed, the risks of detection, seems the major motivation for aggression in social media because persons often aggress under their real names.

Figure 2. 6. Online aggression dependent on anonymity of commenters (random-effects)

Predictions of Table 1, Model 1.

Figure 2. 7. Online aggression dependent on anonymity of commenters (fixed-effects)

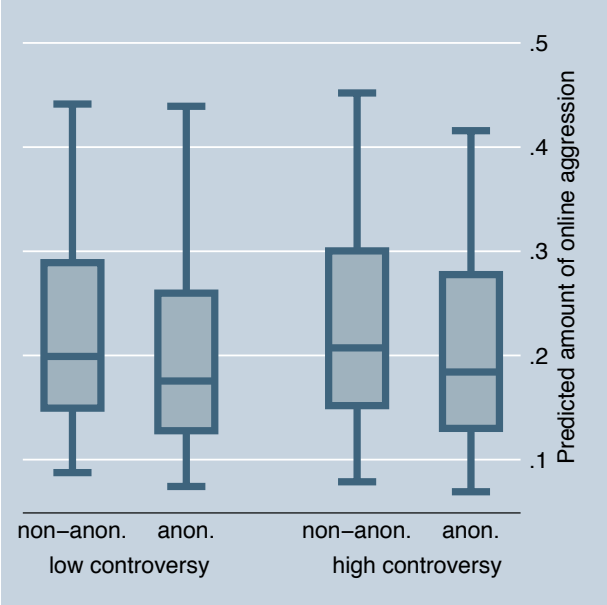
Predictions of Table 2, Model 1.

If norm enforcement is indeed the major motivation for aggression in social media, the highest amount of non-anonymous negative word-of-mouth should be obtained in situations that offer selective incentives and for intrinsically motivated actors. Model 2, in Tables 1 and 2, tests this assumption by introducing interaction effects between the anonymity of commenters and the presence of selective incentives and their intrinsic motivation. The results give preliminary support for Hypotheses 5 and 6. The highest amount of non-anonymous aggression is observed if a petition is accompanied by a highly controversial debate, is connected with a scandal in news media, and if persons are motivated by fairness concerns. By introducing these interaction effects, the main effect of anonymity on online aggression becomes insignificant, and thus suggests that the underlying reasons for non-anonymous aggression can be indeed explained by social norm theory, namely by selective incentives and intrinsic motivation.

Figures 2.2 and 2.8 illustrate the effect for the level of controversy within a debate. In the case of highly controversial topics, individuals clearly prefer to aggress non-anonymously, indicating that selective incentives are present (we code debates as highly controversial if the Herfindahl index is higher than 0.3, and as less controversial if the Herfindahl index is 0.3 or smaller). Figures 2.3 and 2.9 illustrate the effect for the connection with a scandal in news media. Particularly for scandalized topics, the biggest gap arises between the aggression of non-anonymous and anonymous commenters, with the former showing more aggression. Again it supports that scandals offer selective incentives for norm enforcement. Finally, Figures 2.4 and

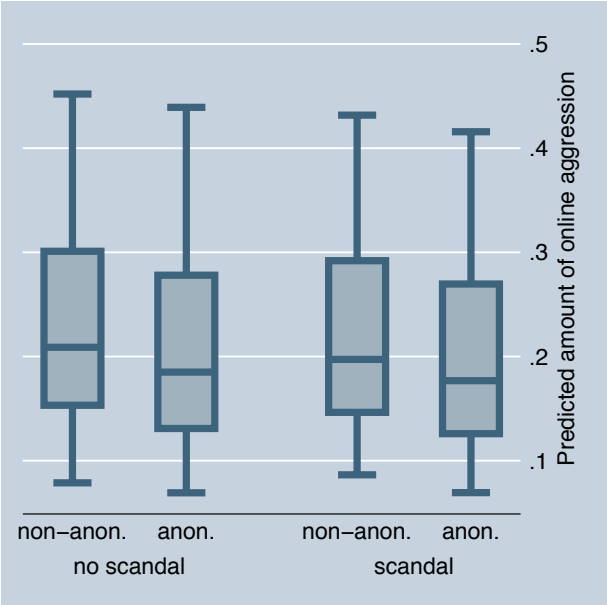
2.5 illustrate the effect for intrinsically motivated individuals. Intrinsically motivated individuals clearly prefer to aggress non-anonymously.

Figure 2. 8. Online aggression dependent on controversy and anonymity (fixed-effects)



Predictions of Table 2, Model 2.

Figure 2. 9. Online aggression dependent on scandal and anonymity (fixed-effects)



Predictions of Table 2, Model 2.

With respect to the control variables, the results show that longer comments and comments submitted earlier in the process of a petition entail a significantly higher amount of aggression. The daily number of protesters has no effect on the amount of aggression, rejecting the assumption that larger petitions attract more negative word-of-mouth. Online aggression

significantly increases for geographically dispersed protests, indicating more general relevance, and for natural persons. Individuals show more online aggression if they live in small villages and cities. We can only speculate about the reasons for this unexpected finding. One explanation is Putnam's (2000) hypothesis that suggests that political participation, and thus also norm enforcement in social media, decrease in large, anonymous regions with a low amount of social capital. Petitions that deal with quality of life entail a significantly lower amount of aggression, whereas petitions that deal with the economy, the media, and environmental or animal welfare entail a significantly higher amount of aggression.

Overall, the random-effects model predicts online aggression by 13%, suggesting that 36% of the variance of aggression can be explained. The fixed-effects model, in which the predictive power is always substantially lower, predicts online aggression by 3%, suggesting that 16% of the variance of aggression can be explained. The predictive power of both models seems rather moderate. One should, however, bear in mind that the predictions are based on objective data, thus implying that common-method biases (and thus systematic-error variance) are absent.

2.7. Discussion

In online firestorms, large amounts of critique, insulting comments, and swearwords against actors of public interest are propagated in social media within hours. This article begins the investigation on this rather new phenomenon by introducing a novel view on online aggression in social media. Relying on social norm theory, we proposed and demonstrated that one major motivation for online aggression in social media is the enforcement of social norms, be it, for example, the struggle for social justice by insulting greedy managers and politicians, or the angst about foreign infiltration by hate speeches against migrants. Norm enforcers punish actors of public interest who cause negative externalities for society or their sub-group by negative word-of-mouth. The technical conditions in social media, such as enhanced visibility and lowered sanctioning costs, have contributed to the expansion of bilateral and multilateral aggressive sanctions which can lead to firestorm-like patterns. Based on this theoretical conceptualization, we also underpinned that online anonymity does not promote online aggression in the context of online firestorms. There are no reasons for anonymity if people want to stand up for higher-order moral principles and if anonymity decreases the effectiveness of sanctions for norm enforcement.

By showing this, we hope to make a number of valuable contributions to the field of online aggression in social media. First, online aggression in a social-political online setting is not primarily an illegitimate and irrational behavior, performed by narcissistic and impulsive actors

with a lack of empathy, social skills and emotional regulation problems acting out of personal revenge (Vandebosch and Van Cleemput 2009, Buckels et al. 2014). Online aggression in social media resembles a practice of sousveillance (Mann and Ferenbock 2013): it accomodates a growing digital civil society that actively uses the available masses of weak ties in social media to publicly enforce social-political norms. Social norm theory offers a theoretical foundation for research on online aggression, which up to now has been largely driven by the absence of theory or psychological interpretations of traditional bullying theory (for example Kokkinos et al. (2014)). Second, it is one of the first studies that has investigated the role of anonymity for online aggression in a social-political online setting by relying on a large dataset that is representative of the proposed digital civil society, i.e., commenters who actively contribute to a wide range of social-political norm enforcement (see also Cho and Kim (2012)). Third, we challenged the popular claim that negative word-of-mouth in social media is mainly caused by commenters' anonymity. In contrast, the results support the idea that non-anonymous aggressive sanctions are more effective. Non-anonymity helps to gain recognition (Douglas and McGarty 2001), increases one's persuasive power (Haines et al. 2014), and mobilizes followers (Andreoni and Petrie 2004). The result is also in line with public voices that observe an increasing social acceptance of non-anonymous digital hate speeches (Connolly 2015).

This study also has practical implications. First, it can be expected that in the future, digital norm enforcement will intensify. The growing digital civil society adapts to the digital environment that transforms interactions. Social media offer great opportunities for individuals who have the intrinsic desire to enforce norms and contribute to the formation of latent interest groups. Second, the regularly demanded abolition of online anonymity and the introduction of real-name policies do not necessarily prevent online aggression in social media. Our view is in line with findings from a natural experiment in South Korea where the enacting of a Real Name Verification Law in 2007 only reduced aggressive comments for a particular user groups, but not for others (Cho and Kim 2012). There is, however, no doubt that the battle over online anonymity will intensify over time, particularly when aggressive norm enforcement by the civil society not only addresses low status, but increasingly high status, actors such as states or corporations.

This study has several limitations that should be kept in mind when interpreting the results. First, the findings are only generalizable to direct, explicitly abusive online aggression but not to indirectly formulated aggression such as cynicism. Also, while we qualitatively checked comments in our large dataset, it was not feasible to identify all comments. The amount of aggression in some comments may be therefore wrongly classified.

Second, in strict terms, the anonymity option of the petition design restricts the generalization of our findings to anonymity hidden from the internet community but not from the petition organizers. However, we consider the transferability to differing anonymity contexts as justified. This is because we do not refer to "true anonymity", but to "relative anonymity", i.e., exploring why spontaneous commenters decide between common options of (non-)anonymity offered for selection by most social media platforms. Achieving true anonymity, in contrast, is difficult anyway: although we recognize that there may be a minority of protesters providing pseudonyms and/or using Tor browsers to hide their identity from petition organizers, and their true anonymity, e.g. to national security agencies, may still not be granted. Consequently, we do not make any inferences on aggressive tendencies by "truly" anonymous commenters because we cannot trace true anonymity and we also expect that the greatest majority of commenters fall back on common (non-)anonymity options.

Third, the results may be not completely transferable to all other types of social media such as criticizing Amazon on Amazon's Facebook fan page. Preexisting norms of cooperation within online petition platforms may lower the expected cost of sanctions. If an aggressive commentator is confronted with a diffuse mass of weak but supportive social ties, he more likely reveals his identity compared to a setting of oppositional ties that could rebuke him, or strong, influential ties that could control inappropriate language.

Fourth, the empirical design does not allow us to draw conclusions with respect to cause-and-effect interpretations. By alternative designs such as most suitably field experiments or intervention studies, it could be analyzed whether the decision to comment (non-)anonymously is indeed driven by social norm deliberations.

Fifth, more information about the protesters and norm violators would be desirable, such as information about their motivation or their socio-demographic characteristics. Exploring whether aggressive protesters differ from non-aggressive protesters on particular dimensions would be of interest here. In regard to aggressors' motivations, another fundamental problematic remains: To what proportion does firestorm-like outrage reflect genuine public opinion? And to what extent does it represent auto-generated propaganda of political (ro-)bots or astroturfers, i.e., fake commenters paid by central coordination units such as political parties? Particularly if public actors increasingly give in to social pressures triggered by firestorms, distinguishing between democratic expression of a legitimate peer-group and a swarm of bots or astroturfers becomes increasingly difficult. Although we perceive the occurrence of bots within our petition data as low (because the lists of signatures finally given to the addressee of

the petition had to include all names and home addresses of signers), this is a challenge that public actors and researchers are likewise confronted with.

While we introduced social norm theory to understand online aggression in social media, many open questions remain. A largely unexplored area is the effectiveness, or offline impact, of digital social norm enforcement. Are there digital accusations that are systematically often ill founded, or mostly justified? Also, beyond knowing that aggressive norm enforcers prefer non-anonymity, how often and under what circumstances do non-anonymous aggressive sanctions indeed help to mobilize other actors and to enforce social norms? Beyond this individual level of analysis, we also recommend focusing on the collective level. A first point is to study, in more detail, the role of selective incentives for (latent) group formation and aggressive acts in social media. Can alternative methods and applications confirm that latent groups aggress more often and mostly non-anonymously? Finally, we did not study the underlying dynamics of online firestorms. Under which circumstances, for example by enforcing which kind of norm and by which framing of sanctions, can online aggressors in social media mobilize other followers within hours?

To conclude, within the increasing penetration of digital media into public life, online aggression has become an effective tool for punishing norm violations and securing public goods. Academia and politics cannot ignore the social-political motivation of an aggressor when investigating online aggression in social media. Also, in the debate on how to legally handle online aggression, underlying social-political motivations must be taken into account in the tightrope walk between securing free expression of opinion and preventing hate speech. And finally, from an ethical perspective, altruistic punishments of norm violations to secure public goods are honorable. However, the question arises whether the aggressive means of punishments as obtained in firestorms are justified.

3. Legitimacy perceptions in online firestorms[6]

Lea Stahel[7], Katja Rost[8]

Abstract

Individuals increasingly use social media to judge the legitimacy of organizations. If certain actions are perceived as unethical, organizations can be hit by large volumes of critical, complaining, and indignant messages on social media, termed online firestorms. This study draws on the legitimacy-as-perception approach to elucidate the micro-level cognitive processes of individuals and their subsequent actions that aggregate to firestorms on the macro level. We argue that individuals use moral heuristics to judge organizational character, procedures, structures, and outcomes as illegitimate. This motivates them to punish organizations aggressively on social media. These micro effects are empirically confirmed by a lexicon-based, manual and automated content analysis of 45,997 comments in a firestorm against a German music rights organization. The study contributes to a more refined understanding of the micro foundations of legitimacy construction in firestorms and to a more sophisticated methodological approach to testing big data from the micro level.

[6] This study is currently in the process of revision and resubmission to the Journal of Business Ethics. It is a strongly expanded version of a working paper published in the Proceedings of the 8th International Conference on Social Media & Society (Toronto, ON, Canada on July 28 - 30, 2017), Article No. 18. ACM: New York, NY, USA, 2017. URL: https://dl.acm.org/citation.cfm?doid=3097286.3097304

[7] Lea Stahel, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland. E-Mail: lea.stahel@uzh.ch
[8] Prof. Dr. Katja Rost, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland.
E-Mail: katja.rost@uzh.ch

3.1. Introduction

Social media increasingly empower ordinary citizens to judge whether organizations are "desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions" (Suchman 1995, p. 574, Etter et al. 2017). They diversify such judgements of legitimacy and so complicate organizations' search for and maintenance of legitimacy (Whelan, Moon and Grant 2013, Etter et al. 2017). The slightest suspicion of an organization violating ethical, social, or moral norms can lead to organizational disruption. This describes a situation where the misbehavior of an organization is initially published or alleged in news media or social media. Subsequently, organizations can be hit by large volumes of critical, complaining, and indignant messages on social media platforms, a phenomenon termed online firestorms (Pfeffer et al. 2014, Johnen et al. 2017). Most organizations are never targeted by firestorms. However, some organizations suffer high-intensity firestorms, in which they are aggressively punished with offensive, threatening, vulgar, and pathologizing comments that spread rapidly and are amplified by news coverage. Such highly intense, aggressive firestorms may harm organizations reputationally, economically, and legally (Pfeffer et al. 2014). They may also polarize perceptions about organizations (Crockett 2017). Some of the most intense firestorms have been triggered by seemingly trivial missteps such as politically or culturally insensitive campaign slogans for the beer brand Bud Light or products offered by Amazon and Zara. Thus, it seems that highly intense, aggressive firestorms are triggered by perceptions of morality rather than how badly organizations actually behave (for initial discussions of morality in firestorms, see Einwiller et al. 2017, or Johnen et al. 2017).

This study innovatively proposes moral heuristics to explain the formation of micro legitimacy judgements and subsequent aggressive actions against organizations in online firestorms. The legitimacy-as-perception theory (Bitektine and Haack 2015, Suddaby et al. 2017) conceptualizes legitimacy as a multilevel social process in which individuals perceive organizations, judge their legitimacy, and act upon these judgments, which ultimately produces macro-level effects on organizations. The theory proposes that individuals use heuristics in general (Kahneman and Frederick 2002, Gigerenzer 2008) to form legitimacy judgements. So far, though, no study drawing on the legitimacy-as-perception approach has discussed the use of moral heuristics in particular. Its theorization and empirical testing thus develops legitimacy research. In this sense, Suddaby et al. (2017) and Bitektine and Haack (2015) argue that exploring the cognitive mechanisms of legitimacy judgments and individuals' consequent actions could shed light on the long-ignored micro foundations of legitimacy. This knowledge is indispensable to fully understanding the construction, maintenance, and demise of

legitimacy. According to Etter et al. (2017), these micro processes should be explored in social media in particular, as social media increasingly influences legitimacy perceptions in the public. However, literature on firestorms has so far only focused on micro communications in static settings such as laboratory experiments and not on volatile online environments (for an exception, see Hewett et al. 2016). This study thus explores how online users in firestorms arrive at judgements of organizations as morally illegitimate and how these judgements motivate them to adopt severe, aggressive punishments.

We integrate moral heuristics (e.g. Bandura 1999, Sunstein 2005; also drawing on Suchman, 1995) into the legitimacy-as-perception approach (Bitektine and Haack 2015, Suddaby et al. 2017). We argue that the organizational disruption motivates individuals not only to be guided more strongly by their personal endorsement of the organization, termed propriety beliefs, but also to use moral heuristics. Specifically, individuals judge organizations as legitimate or illegitimate by up to four moral heuristics. These moral judgements motivate individuals to adopt online sanctions. Online sanctions refer to comments in social media that publicly disapprove of entities that allegedly violate social norms (Rost et al. 2016). We expect that individuals who judge an organization to be morally illegitimate are motivated to adopt particularly severe, aggressive punishments. The aggressive sanctions of individuals then aggregate to a firestorm on the macro level. The study empirically tests the micro effect of moral illegitimacy judgements on aggressive punishment using 45,982 comments from a firestorm against a German music rights organization. We use a lexicon-based, manual and automatic content analysis. This study makes two important contributions. First, by integrating moral heuristics into the legitimacy-as-perception approach and providing empirical evidence, it contributes to a more refined understanding of the formation of cognitive legitimacy judgements (Bitektine and Haack 2015, Suddaby et al. 2017). The second important contribution is its sophisticated methodological approach to micro big data. It enriches the literature on social media communications as a promising field for studying legitimacy construction (Etter et al. 2017) and goes beyond former studies on firestorms by using field data from a high-volume, dynamic, and collective firestorm.

The paper is organized as follows. In the first section, we introduce the legitimacy-as-perception approach and the literature on heuristics and moral heuristics in particular. We then derive a hypothesis about the effect of moral legitimacy judgements on online sanctions in firestorms. In the second section, we discuss our empirical setting, the comment data, and variables. In the third section, we present our results. Finally, we discuss the results, embed them in the existing literature, and suggest directions for future research.

3.2. Legitimacy Perceptions in online firestorms

The legitimacy-as-perception approach (Bitektine and Haack 2015, Suddaby et al. 2017) views legitimacy as a social construct that results from the interaction between individuals and organizations. Individuals are viewed as evaluators, and the socio-cognitive processes underlying their perceptions can be studied (Tost 2011, Bitektine and Haack 2015). The aggregation of perceptions and resulting actions at the micro level supports or challenges the macro legitimacy of organizations. Micro voices such as those found on social media are thereby increasingly influential, an easily accessible data source, and meet the growing interest in the micro foundations of legitimacy in organizational environments (Bitektine and Haack 2015, Etter et al. 2017). Overall, this approach contrasts with others that reduce legitimacy to a mere asset of an entity, an aggregated evaluation of monolithic audiences, or a process (Suddaby et al. 2017).

In this approach, legitimacy is constructed through three components: validity, validity beliefs, and propriety beliefs (Tost 2011, Bitektine and Haack 2015, Suddaby et al. 2017). Validity refers to a socially shared macro opinion about the appropriateness of an organization. It exists as an objective, social fact independent of the opinion of any individual evaluator. On the micro level, evaluators construct legitimacy through judgements, also termed beliefs. To form these beliefs, individuals first need to comprehend organizational practices (Zuckerman 1999, Suddaby et al. 2017). In its most basic form, comprehension means that individuals pay attention to or notice organizations and their practices. Comprehension occurs more likely with increasing organizational visibility (e.g. Rindova, Pollock and Hayward 2006). Validity beliefs refer to individuals' perceptions of whether significant others perceive an organization as legitimate or not, independently of whether the individuals privately endorse that organization as legitimate. Validity beliefs result from validity cues. For example, if an organizational practice has a noticeable influence on the opinions and actions of surrounding actors, this is a cue that this practice is legitimate. However, in situations of turmoil, individuals more likely trust their propriety beliefs than their validity beliefs. Propriety beliefs describe individuals' private endorsement of organizations and their behaviors. In situations where organizations struggle to maintain legitimacy, validity is weakened because there is no societal consensus on the organizations' legitimacy. Thus, validity beliefs are not clear-cut. Accordingly, individuals' perceptions are more strongly guided by their propriety beliefs. The legitimacy-as-perception approach in particular recognizes such heterogeneous judgements of individual evaluators, which emerge particularly in situations of turmoil (Desai 2011, Bitektine and Haack 2015).

The formation of legitimacy judgements may be explained by socio-cognitive heuristics. These are distinctive mental operations that serve as anchors for legitimacy judgments (Bitektine and Haack 2015, Suddaby et al. 2017). If a target is noticed, but individuals seek to minimize their cognitive effort when judging it, they will choose heuristic shortcuts instead of carefully evaluating all the information or statistics (Kahneman and Frederick 2002). The mechanism underlying heuristics is attribute substitution: if individuals have to make a judgement about a target attribute that is cognitively too complex, they instead substitute the target attribute with a heuristic attribute which comes more readily to mind (Kahneman and Frederick 2002: 4). For example, if individuals face a difficult, novel problem, they search for a more familiar, similar problem and transfer its solution to the more difficult problem. They select from a long list of heuristics, and each of these induces its own systematic bias and errors (McPherson, Smith-Lovin and Cook 2001, Gigerenzer 2008). For example, if individuals are able to assign organizations to a well-defined category, their positive and negative evaluation of the organization will be more clear-cut and less neutral (Vergne 2012).

Applying the legitimacy-as-perception approach to firestorms, we argue that the legitimacy of organizations is constructed in a social process between online users in social media. These individuals perceive, judge, and act through digital platforms such as social networking sites, online commentary sections, and online petitions. Their social environment is characterized by diverse collectives composed of individuals on particular online platforms and by news media as a powerful authority. Information technologies enable an unlimited number of individuals to notice, or comprehend, organizational disruptions and subsequently form beliefs. This is because information technologies allow any unethical organizational behaviors that occurred anywhere at any time to be easily unearthed, made visible, and publicly alleged (Pfeffer et al. 2014). Such information is persistently stored and can be replicated, shared, and made accessible to unlimited audiences by news media and social media (Marwick and Boyd 2011). Cues such as the numbers of comments opposing an organization and online likes help to form consequent validity beliefs. In situations of organizational disruptions, however, validity will be weakened. This is because these situations reflect organizational turmoil where the – so far potentially unquestioned – collective legitimacy of the organization and its practices are suddenly questioned and any related validity beliefs thus less clear-cut. Individuals will thus be guided by their personal propriety beliefs more than usual.

The digital context not only helps individuals to notice organizational disruptions; it is also an environment in which individuals readily turn to socio-cognitive heuristics to form propriety beliefs. Online, individuals are tempted to minimize their cognitive effort even more than usual.

This is because in the media-saturated, information-rich context of the Internet, individuals experience an overload of information while their attention is limited, so information consumption is constrained (Fairchild 2007). In this volatile online context, information diffuses rapidly and far. One example is Twitter, which obliges communication to be short and quick. In addition, the opportunity window to join a debate about any organizational disruption is small. Commonly, such online debates are in full swing before the details and background of the particular organizational disruption can even be published. Overall, this invites quick reactions rather than long inquiries, which tempts individuals to take shortcuts through heuristics. Lastly, while general information is abundant online, the specific, individualizing information that commonly inhibits the use of heuristics is scarce. For example, the digital context generally lacks nonverbal cues such as facial expression or physical size. This reduces the impact of empathy and authority typically expressed by body language, dress, and social setting, which encourages the use of heuristics and the projection of enemy images (Kiesler et al. 1984, Suler 2004, Mehari et al. 2014).

Suchman's (1995) classification of moral legitimacy into four dimensions, namely organizational character, procedures, structures, and outcomes, can be used to describe the socio-cognitive heuristics that individuals may use when online to categorize and stereotype organizations. Moral legitimacy is a heuristic commonly used as a central benchmark when organizational behavior is discussed in public, political-normative arenas (Palazzo and Scherer 2006). Sunstein (2005) argued that moral heuristics are pervasively used in political and legal contexts to reduce the highly complex problems commonly debated in these areas. Similarly, organizational disruptions are commonly published, alleged, and debated in political-normative online arenas, so moral heuristics are likely to be commonly used. They allow individuals to make a judgement about an organizational disruption while avoiding immersion in the elusive depths of the disruption and ignoring the overwhelming information online. Inspired by Suchman (1995), we argue that individuals judge an organizational disruption as morally illegitimate, or morally legitimate. If they view an organizational disruption as immoral, they will form propriety beliefs that substitute the potentially complex attribute central to the disruption with more easily accessible, negatively connotated attributes about organizational character, procedures, structures, and outcomes. Accordingly, individuals may substitute an organization with another social category or character, such as a negatively evaluated industry, network, government, or world order independent of any valid comparability. Organizations are thus reduced to exchangeable representatives of the category instead of being recognized as unique actors with differentiating characteristics. Empirical studies have confirmed the

occurrence of such spillover (Jonsson, Greve and Fujiwara-Greve 2009, Haack, Pfarrer and Scherer 2014). Individuals may also use heuristics to judge organizational procedures as morally illegitimate, for example how organizations arrive at knowledge (e.g. pseudoscience), how they produce, trade, and organize working conditions (e.g. child labor or unequal waves), how they treat customers and design marketing campaigns, and how their management decisions are made (e.g. lack of transparency). Heuristics can also be used to judge organizational structures as illegitimate. When this happens, organizational structures are sweepingly accused of being incompatible with the current, socially accepted structural environment. For example, individuals substitute organizational structures with seemingly similar, but condemned, structures typically located in different times and places, such as Taylorism, oppressive regimes, and the Middle Ages. Finally, heuristics may be used to judge organizational outcomes. For example, individuals may consult their memories about similarly negative experiences with the organization. This includes past failures and misbehaviors.

Individuals who use heuristics to judge organizations as morally illegitimate may well adopt aggressive online sanctions. Generally, online sanctions are common actions adopted by individuals in firestorms to put costs on organizations, to induce conformity, and ultimately, to secure public goods (Rost et al. 2016). Online sanctions can differ in severity (Douglas Creed, Hudson, Okhuysen et al. 2014, Antonetti and Maklan 2016, Crockett 2017). For example, individuals may publicly disapprove organizations but maintain a neutral tone in their online comments. We call this sober disapproval. Their commenting may also include emotional and scandalizing language such as emotional shaming. However, the most severe sanction is aggressive comments. They punish organizations by offensive, vulgar, pathologizing, and threatening language. We argue that using moral heuristics to judge organizational disruptions as morally illegitimate will encourage individuals to punish organizations aggressively. This is because moral heuristics introduce ideological justifications and associated emotions that reduce individuals' external and internal sanctioning costs (Bandura 1999, Crandall and Eshleman 2003, Haslam 2006). Moral heuristics accordingly lead individuals to feel morally obliged to secure moral norms and to perceive an increased need for effective sanctioning.

Aggressive punishments are particularly effective because they can be very costly for organizations. They are the most visible as their highly arousing emotive content has the highest potential to spread in social media. They are also most convincing because they signal the willingness of individuals to accept high sanctioning costs (Stieglitz and Dang-Xuan 2013). In summary, the justifications introduced by moral heuristics change classic cost-benefit concerns and encourage the adoption of otherwise costly, aggressive behavior. For example, it has been

shown that individuals whose expectations have been confounded by organizations feel strong moral outrage (Antonetti and Maklan 2016), which motivates boycotts (Lindenmeier, Schleer and Pricl 2012), revenge (Barclay, Whiteside and Aquino 2014), and whistleblowing (Jones, Spraakman and Sánchez-Rodríguez 2014). We therefore hypothesize that *individuals relying on moral heuristics to judge organizations as morally illegitimate in terms of character, procedures, structures, and outcomes will adopt aggressive online punishments against these organizations more strongly than those who do not judge organizations as morally illegitimate.*

The aggregation of aggressive online sanctions results in the macro phenomenon discernible as a firestorm. Firestorms may challenge the macro validity of organizations. Judgments expressed by individuals aggregate to construct, preserve, or challenge the macro validity of organizations and so drive institutional change (Bitektine and Haack 2015). Suddaby et al. (2017) propose both economic theory (Kuran 1997) and system justification theory (Jost, Banaji and Nosek 2004) to explain the micro–macro translation. Both theories suggest that if legitimacy judgements can be observed on the micro level, the validity of these judgements also increases on the macro level. For example, when aggression is expressed by a majority of individuals, incoming individuals are encouraged to conform to this aggressive norm. This is because validity beliefs about how the social environment thinks and acts about the organizational disruption are changed. In this case, individuals expressing judgements of moral illegitimacy and subsequently adopting aggressive punishment expect to be cheered by others. The more visible aggressive sanctions are, the stronger is their impact on the macro validity of the organization. Visibility increases if platforms possess powerful diffusion mechanisms such as liking, sharing, or retweeting; these amplify information sharing, costless feedback, and habitual outrage expression (Stieglitz and Dang-Xuan 2013, Crockett 2017). Visibility also increases if news media cover the firestorm (Bitektine and Haack 2015). Changes to macro validity may then impact organizations reputationally, economically, and legally.

3.3. Data and Method

3.3.1. Empirical setting and data

We use 45,997 online comments submitted to a German online petition (https://www.openpetition.de/petition/online/gegen-die-tarifreform-2013-gema-verliert-augenmass) against the *Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte* (Society for Musical Performing and Mechanical Reproduction Rights: GEMA). For permission of using the data, see document 1 in the Appendix. The petition was organized to protest against a licensing fee scheme planned in 2012 by GEMA. Diverse actors

such as club managers feared increases in fees. An event promoter launched the online petition in April 2012. In October 2012, 183 days later, the petition was closed with 305,122 signatures. The petition signatures were presented to the Federal Minister of Justice of Germany in December 2012. A week later, GEMA announced that it would delay the implementation of the tariff reform to allow further negotiations.

This data is optimal for testing our hypothesis. First, the petition platform explicitly invites all petition signers to add comments describing their personal motivation for signing. Signers express their legitimacy judgements and their sanction verbally. Furthermore, the GEMA petition's content and the online platform setting is representative for firestorms. The GEMA petition shares the central attributes that differentiate firestorms from other forms of online communication (Johnen et al. 2017). The petition has a "moral concern": many individuals judge GEMA to be morally illegitimate. There is "minimal consensus": most individuals condemn GEMA. Further, the petition is "hostile": individuals are frequently aggressive against GEMA. It is also "disproportionate": the petition attracts unusually large numbers of participants all over Germany. Finally, the petition is "volatile": it exhibits several daily peaks, including one up to 2366 comments a day, as descriptive statistics of the present data set show. The petition was collected on a well-known social media platform where the petition signers cooperated to reach a common goal, the achievement of the petition objective, and publicly produced and consumed comments in doing so (Treem et al. 2016). The resulting chronological list of comments reflects one of many possible firestorm structures. Similar firestorms may occur on platforms such as Facebook, Twitter, Youtube, and news commentary sections. The absence from the petition of any option for commenting back and forth also complies with the typically unilateral orientation of firestorms: firestorm participants commonly interact little with each other but predominantly sanction external entities.

We complement this data with two additional data sources. The first is metadata on commenting individuals, including their degree of anonymity, time of comment submission, and postcode. The platform owner provided the comments and metadata to the authors on request. The second data source is a group of news media articles on GEMA, drawn from the Lexisnexis online media database. None of the data requires research ethics board approval. This is because all data is publicly accessible and was provided to the authors in an anonymized form. Thus, no individual user can be identified.

3.3.2. Method

This section presents the process by which we coded our variables; the specific operationalization of these variables is presented in the subsequent section. To test our hypotheses, we processed the qualitative comments into quantitative data. We used a lexicon-based, manual and automated context analysis (Grimmer and Stewart 2013). The first step was manual content analysis, also termed human hand coding. For this, 5% of the randomized comment corpus was selected (around 2000 comments). One coder read and classified all comments in this subset into the pre-defined categories (for example, into aggressive punishment). This process created coding rules. Coding rules are intended to guide classification of comments into categories. Coding rules were refined in a circular process of classifying and discussing with a second coder to ensure valid, unambiguous, and mutual exclusive categories. To validate the rules for classification, both coders classified 500 comments (1% of the randomized corpus) into the categories. The resulting interrater-reliability of Cohen's Kappa was substantial to excellent. For aggressive punishment, it was .88, for emotional shaming, it was .87, and across all four moral heuristics, it was ø = .77 (minimally .70).

The second step was dictionary building. Dictionaries are basically word lists; dictionary methods use the frequency of these words to automatically classify a document, for example each comment in a comment corpus, into a category (Grimmer and Stewart 2013). One coder selected all of the comments that were classified in each of the predetermined set of categories. The coder then identified all words, word combinations, and punctuation marks that discursively embodied the category. For instance, the expression "idiots" was classified as *aggressive punishment (*all of the following text examples were translated from the original German text to English by the authors). All the expressions identified were transferred to one dictionary for each category. However, the expressions needed to be transformed into linguistic forms that are not misclassified by automatic classification. Misclassification occurs, for example, if too many false positives are counted; instances of expressions used with other meanings than intended. We used two tools for the linguistic transformation. The first tool is the part-of-speech tagging tool TreeTagger. TreeTagger differentiates word searching and simplifies vocabulary. It allows lemmatization, that is, using the basic form of words to detect an item in all its possible forms (e.g., a search for 'good' also searches for 'better' and 'best'). It allows the search of expressions to be manually restricted or expanded to any desired endings, inflexions, and conjugations, and consecutive combinations of words to be searched for despite having irrelevant words in between. However, to specify the linguistic forms that optimally

search for expressions in the desired meanings but not in others, coders need to know the linguistic contexts in which expressions are used within a corpus. To this aim, we used the online tool Sketch Engine. It provides an overview of all the sentences in which a selected expression occurs.

The third step is automated classification using Python. This displays a count value for each category for each comment of the whole corpus. This value describes how many expressions from each dictionary are found in each comment. In our corpus, the automated classification reliably replicated manual coding: Cohen's Kappas between automated and manual coding for 1% of the randomized material is substantial to excellent. For aggressive punishment, it is .89, for emotional shaming, it is .92, and for moral heuristics, it is ø = .82 (minimally .63).

3.3.2. Measurements

Aggressive punishment measures the sum of offenses (e.g. "liar", "fraudsters"), vulgar speech (e.g. "fuck", "puke"), ascribed pathology (e.g. "GEMA is insane"), and threat of violence or abolishment (e.g. "GEMA workers should be whipped in public", "Your institution will burn") in each comment. These four subcategories emerge from established measurements of online aggression. This dictionary contains 346 expressions. The variable is over-dispersed: 86% of the commenters use non-aggressive expressions, 11 % use one aggressive expression, and only 3% use two to six aggressive expressions. The GEMA petition case thus represents a low-intensity firestorm in percentage but a high-intensity firestorm in absolute numbers, with 6608 aggressive comments overall.

In our theory, we differentiate aggressive punishments from the less severe sanctions of emotional shaming and sober disapproval. For robust results, additional analyses should find that individuals using moral heuristics do not use less severe sanctions.

Emotional shaming measures the sum of emotional display (e.g. NOT, !!, ?!?, ☹), emotional language (e.g. "angry", "afraid"), and scandalization (e.g. "outrageous", "shame on you!") in each comment. The dictionary contains 76 expressions. Of all comments, 24% include emotional shaming.

Sober disapproval is a binary variable created through a process of exclusion. First, each comment is coded according to whether it contains emotional shaming and/or aggressive punishments. In our study, 64% include neither. These comments are classified as sober disapproval because they represent the public posting of disapproval towards entities or their behaviors. Examples are: "I am against the reform" or "I hope that the only pub in my village will not have to close".

The following four count variables capture the moral heuristics by which individuals classify GEMA as morally illegitimate.

Illegitimate character measures the sum of expressions that stereotype GEMA as a morally illegitimate character in each comment. In these cases, individuals follow one of three strategies. Firstly, they stereotype GEMA directly, for example, as representative of the "bad" music industry, of a monopolistic organization, or of the neoliberal world. One such comment is "I am against the enrichment of monopolists". Secondly, individuals indirectly stereotype people working for GEMA as, for example, "civil servants". Finally, individuals stereotype GEMA by drawing analogies to other stereotyped entities such as banks. One such comment is "Oil companies… GEZ [fee collection service of Germany's public broadcasting institutions or *Gebühreneinzugszentrale der öffentlich-rechtlichen Rundfunkanstalten in der Bundesrepublik Deutschland*; author's note] … GEMA… – none of them can get enough". The dictionary contains 51 expressions. Of all commentators, 5% (2278 individuals) in our study stereotyped GEMA as a morally illegitimate character.

Illegitimate procedures measures the sum of expressions that judge GEMA's procedures as morally illegitimate in each comment. GEMA is predominantly accused of illegitimate business and customer interactions and unfair procedures. Individuals criticize, for example, GEMA's mismanagement ("inefficient"; "not transparent"), reckless methods ("egoistic"; "arbitrary"), disrespect or exploitative treatment of clients ("Without any willingness for compromise"), and unfair acting ("This reform is unfair"). The dictionary contains 329 expressions. In our study, 19% (9135 individuals) judged GEMA's procedures to be morally illegitimate.

Illegitimate structures measures the sum of expressions that judge GEMA's structures as not complying with current, morally legitimate structural environments in each comment. For example, individuals equate GEMA' structures with commonly condemned political systems such as dictatorship regimes and unpopular historical ages (e.g. "We do not live in medieval times"). Or, they ascribe an obsolete, outdated identity that has no reason for existence (e.g. "GEMA is a useless and antiquated institution"). The dictionary contains 107 expressions. Some 2.4% (1070 individuals) judged GEMA's structures to be morally illegitimate.

Illegitimate outcomes measures the sum of expressions that judge GEMA to have a history of morally illegitimate behaviors in each comment. For example, individuals describe the present behavior as typical of a history of misconduct and maladministration. Or they express prolonged dissatisfaction with GEMA and emphasize the imperative for change. Comments include "Put a stop to this game finally!" and "Now, it is getting too much with GEMA ... I

have been annoyed for years". The dictionary contains 91 expressions. Some 7,3% (3364 individuals) judged GEMA's outcomes to be morally illegitimate.

3.3.3. Control variables

We control for variables that may influence individuals' adoption of online sanctions.

Sophistication of language is a metric variable that measures the linguistic complexity of comments. For each comment, we use the Flesch-Kincaid-grade-level formula (Kincaid, Fishburne, Rogers et al. 1975) to calculate a score of how difficult it is to read: [0.39*(total words/ total sentences) + 11.8*(total syllables/ total words) - 15.59]. This formula has already been applied to measure readability of online reviews (Ghose and Ipeirotis 2011). The higher the score, the longer the sentences and words, and the more difficult it is to read. The measure is only reliable for longer texts. Therefore, we assign the lowest score to the 27% of comments that include less than seven words. Higher scores suggest higher literacy skills. This may decrease the attractiveness of aggressive punishment because the cognitive costs needed to formulate reflected argumentation is lower for such individuals.

We use two proxies to control for individuals' arousal. *Arousal through spelling mistakes* counts the number of spelling mistakes in a comment. If people are more aroused, such as through intense eye contact or unpleasant music, they are less able to process information. This reduces their performance in memory tasks and word-spelling (Conty, Russo, Loehr et al. 2010). We expect that more strongly aroused individuals punish more aggressively because their cognitive costs for more reflected argumentation are higher. *Arousal level of words* measures the average arousal of words that individuals use in comments. Research (Storbeck and Clore 2008) suggests that if individuals are more aroused, arousing words are more accessible and thus more often used. It is measured by the German adaptation of the dictionary 'Affective Norms for English Words' or ANEW (Schmidtke, Schröder, Jacobs et al. 2014). This German translation of the ANEW material lists 2902 German words and rating of how arousing they are. If one or more of the listed words appear in a comment, we assign the mean of the respective arousal rating(s) to the comment.

Anonymity indicates whether individuals chose to allow either their real names and residence to be published beside their comments (0 = non-anonymous) or only their postal codes (1 = anonymous). A positive by-product of this procedure is that it minimizes the risk of social bots that could distort our results. With 41% of anonymous individuals, the petition is about average for user anonymity on social media platforms: for example, anonymity is higher on YouTube

and lower on Facebook. Non-anonymity may increase aggressiveness in social-political commenting (Rost et al. 2016).

Self-reported economic dependency measures whether or not individuals self-report working in or owning a business that is threatened by the GEMA reform (e.g. "As a club owner I could go bankrupt" or "I could lose my job as DJ"). The dictionary contains 124 expressions. Of all individuals, 4.8% self-report being economically dependent. Self-reported economic dependency may either increase aggressive punishment, because the GEMA reform puts higher cost on these individuals, or it may decrease aggressive punishment, because publicly expressing personal information increases the risk of external sanctions.

Urbanity measures the size of the city or village in which individuals live (by the approximate number of inhabitants per 100,000). Higher values indicate more urban places. As a socio-structural variable, it influences life cultures and associated sanction preferences.

We also control for the tone of the daily news media coverage on GEMA. News coverage is a 'validity cue' that signals macro validity and thus influences the expected sanctioning costs of micro evaluators (Bitektine and Haack 2015, Einwiller et al. 2017). Using the LexisNexis media database, we searched for the term GEMA in the title, lead, or main text of all German-speaking print and online media articles on GEMA during the petition period. Then, we manually selected articles that are associated with the studied GEMA reform. In the resulting 789 articles, we first measure *Negative media coverage*. One coder accordingly determined whether each article reports predominantly negatively on GEMA: whether it gives more attention to contra than pro viewpoints on the reform. In total, 479 negative articles were identified and summed. Similarly, for *Balanced media coverage*, the coder determined whether each article reports in a predominantly balanced way on the GEMA reform: whether it gives the same attention to pro and contra viewpoints on the reform or contains merely descriptive news without taking up any arguments. In total, 234 balanced articles were identified and summed. Cohen's Kappa interrater reliability with a second coder for 10% of all articles was ø = .94.

We control for two other validity cues, the number and content of previous comments. *Previous comments (total)* measures how many comments (per 1000) were submitted prior to each individual comment. *Previous aggressive punishment* measures the number of aggressive punishers in the ten previous comments. On this petition platform, we expect effects to be weak, as the platform design requires individuals to read and submit comments on separate pages, and this increases the costs of detecting validity cues.

Number of words measures the sum of words in a comment (per 100 words). More overall words increase the number of aggressive expressions.

For descriptive statistics and correlations of variables, see Table 6 in the Appendix. For the statistical analyses, we apply regressions on the individual comment level. Negative binomial regressions are selected for predicting aggressive punishment and for emotional shaming. This suits their highly dispersed, count-data structure. Logistic regression is selected for predicting sober disapproval. We group the dataset for each of the 183 petition days as comments are expected to correlate within each petition day. The more general, robust population averaged estimators of generalized estimating equation (GEE) regressions are applied for all models. They are an alternative to fixed and random effects models. GEE models suit the present data best as they address unmeasured dependence in clustered data and outcomes.

3.4. Results

Table 3 reports the findings. The results in Model I confirm that individuals who rely on moral heuristics and judge GEMA as morally illegitimate indeed adopt aggressive punishments more strongly. The effect of morally illegitimate character on aggressive punishments is stronger by far than those of the remaining moral dimensions. The regression coefficients show that one expression of an illegitimate character leads to 0.5 additional aggressive expressions, while this value is 0.1 for procedures, 0.2 for structures, and 0.1 for outcomes. It indicates that moral mental shortcuts punish organizations much more strongly if individuals reduce organization to a simplified, negative character instead of considering its procedures, structures, or outcomes. The control variables indicate that that individuals adopt aggressive punishment more strongly the less sophisticated their language, they more aroused they are, if they do not self-report being economically dependent on the GEMA reform, if they are non-anonymous, the less urban they live, the more negative and the fewer balanced news media articles about the GEMA reform are published on the particular day, and the more words the comment contains. On this petition platform, the number and content of previous comments did not impact aggressive punishments.

Table 3. Regression effects of moral heuristics on online sanctions

	Aggressive punishments (negative binomial regression)				Emotional shaming (negative binomial regression)				Sober disapproval (logistic regression)			
	Coeff.	Std.Err.	z		Coeff.	Std.Err.	z		Coeff.	Std.Err.	z	
Moral heuristics												
Illegitimate character	0.53	0.03	15.53	***	0.05	0.04	1.27		-0.36	0.04	-9.18	***
Illegitimate procedures	0.07	0.02	3.61	***	0.06	0.02	3.61	***	-0.08	0.02	-4.32	***
Illegitimate structures	0.18	0.06	2.99	**	-0.37	0.07	-5.44	***	0.17	0.06	2.84	**
Illegitimate outcomes	0.08	0.03	2.61	**	0.04	0.03	1.68		-0.10	0.03	-3.89	***
Control variables												
Sophistication of language	-0.05	0.00	-22.48	***	-0.05	0.00	-27.95	***	0.08	0.00	37.85	***
Arousal through spelling mistakes	0.10	0.00	22.00	***	0.02	0.00	5.31	***	-0.07	0.01	-12.06	***
Arousal level of words	0.11	0.02	5.79	***	-0.08	0.02	-5.37	***	0.01	0.02	0.94	
Anonymity	-0.18	0.03	-7.13	***	-0.01	0.02	-0.47		0.11	0.02	5.43	***
Self-reported economic dependency	-0.70	0.08	-9.19	***	-0.11	0.05	-2.18	*	0.37	0.05	7.31	***
Urbanity of residence	-0.05	0.01	-3.87	***	-0.05	0.01	-4.50	***	0.06	0.01	5.81	***
Negative media coverage	0.01	0.00	2.42	*	0.00	0.00	0.19		0.00	0.00	-1.06	
Balanced media coverage	-0.01	0.01	-2.35	*	0.00	0.01	0.56		0.00	0.00	0.42	
Previous comments (total)	0.00	0.00	0.19		0.00	0.00	0.08		0.00	0.00	0.19	
Previous aggressive commenting	0.01	0.01	0.74		0.00	0.01	0.05		0.00	0.01	-0.72	
Number of words	0.30	0.05	5.83	***	1.02	0.04	24.91	***	-1.09	0.06	-19.52	***
Constant	-1.90	0.08	-22.89	***	-0.74	0.07	-10.51	***	0.20	0.07	2.85	*
Number of observations			44,173				44,173				44,173	
Wald chi2			1751.39	***			1418.96	***			1882.47	***
Number of clusters (days)			179				179				179	

* Significant at $p < .05$; ** significant at $p < .01$; *** significant at $p < .001$

We add two robustness tests in Table 3. Model II predicts the amount of emotional shaming and Model III predicts the likelihood of sober disapproval. For robust findings on aggressive punishments in Model I, no effects should be observed for emotional shaming and opposite-effect directions for sober disapproval. In general, this is confirmed by the data. How strongly individuals adopt emotional shaming is not affected by how strongly they judge GEMA's character and outcomes to be morally illegitimate. Also, individuals are more likely to adopt sober disapproval if they do not judge GEMA's character, procedures, or outcomes as morally illegitimate. We also find surprising opposite effects: individuals who judge GEMA's structures as morally illegitimate are not only more likely to adopt aggressive punishment but also sober disapproval. Simultaneously, they are less likely to adopt emotional shaming. Additionally, individuals who judge GEMA's procedures to be morally illegitimate are not only more likely to adopt aggressive punishment but also emotional shaming.

3.5. Discussion

Our results show that individuals more strongly adopt severe, aggressive punishments if they judge entities' characters, procedures, structures, and outcomes to be morally illegitimate. Moral heuristics thus introduce justifications to consider aggressive punishments as 'the right thing to do' and, accordingly, change classic cost-benefit concerns. This encourages the adoption of adopt otherwise costly, aggressive behavior. Our study advances legitimacy research by shedding light on its long ignored cognitive micro foundations. The results confirm the presence of heuristics in micro legitimacy construction processes (Kahneman and Frederick 2002, Gigerenzer 2008), the dependence of actions upon heuristic-based judgements (Tost 2011, Bitektine and Haack 2015, Suddaby et al. 2017), and the effects of moral justifications on aggressive behaviors (Bandura 1999, Crandall and Eshleman 2003, Haslam 2006). Overall, the results suggest that morality and its varying manifestations play a critical role in aggressive firestorms. This particularly applies to stereotyping an organization as an illegitimate character. Ultimately, organizations may profit from this study. Organizations find difficulty in addressing the diversity created through social media and the dynamics involved in firestorms. However, aggressive firestorms seem to be driven by perceived moral illegitimacy, so organizations could strengthen or weaken such judgements through organizational behaviors and communication, particularly through social media. This may be critical, as moral heuristics commonly induce systematic errors in reasoning which may harm a debate (Sunstein 2005). Further, organizations

are suggested to particularly look out for the four moral legitimacy dimensions in social media discourses to anticipate aggressive firestorms.

The second important contribution is the sophisticated methodological approach to analyzing the data from a firestorm. The exceptional use of field data from firestorms increases the external validity of our results and goes beyond former empirical designs using fictitious firestorm data. The lexicon-based manual and automatic content analysis is a so-far underused method to study legitimacy; it thus meets the call to expand the methodological toolkit for studying this complex phenomenon (e.g. by Suddaby et al. 2017). Our approach performs well in context sensitivity and allows deep exploration of data and the simultaneous processing of a large text corpus instead of blindly following domain-specific, pre-coded word lists. Methodologically and datawise, this study thus enriches the literature that emphasizes micro communications on social media as a promising new data source for studying organizational legitimacy construction (Etter et al. 2017). Beyond, social media discourses deserve attention as they have a growing influence on macro legitimacy (Etter et al. 2017) and are thus a powerful tool for enforcing norms in the digital age (Rost et al. 2016, Crockett 2017).

The study opens important avenues for future research. First, micro–macro interactions in legitimacy construction invite empirical investigation. For example, how do moral judgements mediate between socio-structural macro conditions and individuals' choices of online sanction? Sociodemographic data on individuals collected by surveys and linked with social media data could help to explain why moral judgements about structures seem to polarize the sanctions adopted; they encourage both aggression and sober disapproval. A second empirical avenue is to validate the proposed effects in a larger sample of firestorms. We suggest testing the effects across diverse targeted organizations, societal fields, cultural settings, and platforms. This would add to the case in the present study, an isolated firestorm on a single online platform. The use of methodologies that both efficiently process big data sets and effectively detect dynamical legitimacy constructs such as complex linguistic analysis or machine learning (Etter et al. 2017) is promising. Third, controlled lab experiments may test the causal link between moral legitimacy judgements and online sanctions in micro processes. Social and cognitive suppressing factors could also be manipulated in such studies (Suddaby et al. 2017).

In the global village, for good or ill, the social, hierarchical, and geographical barriers to detecting and sanctioning organizational behaviors deemed illegitimate have fallen. The resulting critical online debates confront organizations with ordinary citizens' norms, expectations, and normative demands that could be largely neglected in the past. This

diversification of judgements complicates the search for and maintenance of legitimacy. By exploring the moral heuristics underlying micro judgements in social media, we hope this study sheds early light on legitimacy construction in the progressing digital age.

4. "Dirty journalists, all liars!" - A social identity explanation for why journalists are attacked by audiences[9]

Lea Stahel[10]

Abstract

Journalists are offended, threatened, and disparaged by audiences through digital and analogue channels. This negatively impacts journalists and democracy. It remains unclear which journalists are particularly frequently attacked and why. This study draws on five central conditions that a politicized social identity approach suggests increase threats to groups' social identity and power. Journalists should be more frequently attacked if they are evaluative, publish on political topics, have a local focus, are powerful, and belong to outgroups that are hard to distinguish from the ingroup. This is because their potential to threaten the social identity and power of groups is greater under these conditions; they are more likely to 'mobilize' group members with a pre-existing tendency to feel threatened. A Swiss online survey on 530 journalists confirms all five hypotheses. Results support a perspective on aggression against highly exposed professions in digital societies that is inspired by sociological social psychology.

[9] This study is submitted to the European Sociological Review.

[10] Lea Stahel, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland. E-Mail: lea.stahel@uzh.ch

4.1. Introduction

With such a name, you cannot be a journalist. Get me a real journalist.

Dirty journalists, all liars!

You have wife and kids and we know where you live.

Nowadays public professions such as journalists, politicians, celebrities, and senior executives are more visible and accessible than ever before. To attract such aggressive comments, journalists in particular need not live in a country where journalistic freedom and autonomy is suppressed – these attacks are also received by journalists in democratic, Western countries. The rise of communication technologies plays its part in this regard. The Internet allows news organizations to engage and inform audiences all over the world (Lee 2015) and journalists to gain information and networks and promote themselves through social media (Hedman and Djerf-Pierre 2013). They also enable readers, spectators, and listeners to provide instant, direct and global feedback to journalists. In the best case, this enables fruitful exchange and enhances journalistic quality (Ksiazek et al. 2015).

The reality, however, often fails to attain such noble ends. The opportunities for audiences to scrutinize, devalue, and discredit journalists and their published output have become wide open. Journalists are frequently attacked by their audience: they receive vulgar, pathologizing, inappropriately generalizing, disparaging, offensive, and threatening feedback about themselves and their output. They are attacked digitally through news commentary sections, email, social networking sites, and – still – through letters and face-to-face attacks. This phenomenon demands public and scientific attention. Although aggression has always been part of public discourse, it "today has more outlets, can be highly public, and travels and spreads faster; its impact can be greater, and strategies for dealing with it are still being tested" (Meltzer 2015: 86). Initial and highly informative reports from Germany (Preuss, Tetzlaff and Zick 2017), Sweden (Löfgren Nilsson and Örnebring 2016), and Hungary (Tofalvy 2017) show that half to three quarters of journalists surveyed were attacked at least once within the year prior to data collection. They also show that aggression impacts journalists profoundly negatively. They feel fear and self-censor. Aggression also harms democratic outcomes: in experiments, audiences exposed to it polarize their opinions, distrust media, and think of society as more divided (e.g. Anderson et al. 2016).

Two so far open questions are which journalists are most frequently attacked by audiences, and why. Qualitative studies on online hate against female journalists suggest gender differences (e.g. Hardaker and McGlashan 2016), while purely descriptive surveys initially find differences between media organizations and topics (Löfgren Nilsson and Örnebring 2016, Preuss et al. 2017). These studies are commonly atheoretical, descriptive, or case-oriented. No theoretically driven, systematic, multivariate empirical approach has yet shed light on this phenomenon. This study fills this gap by asking which characteristics of journalists influence the frequency with which they are attacked by audiences, and why.

A comprehensive approach to answering this question might examine journalists' potential to threaten the power of social groups. Power includes the control that groups possess over their own and outgroups' outcomes and the possession of immaterial resources such as information, expert knowledge, status, and reputation (Simon and Klandermans 2001). Journalists may influence groups' power in various ways. For example, journalists scratching at popular politicians' images may also embarrass their electoral groups. Journalists tackling local corruption may disrupt deep-seated social group ties. Such threats may be amplified if published by journalists who are well-known because they commonly address a large audience. Members of social groups may accordingly be enraged and intervene by attacking the journalists they hold responsible for their anger. In a society pervaded by digital technologies, this is easier than ever before, because the reduced costs of communication and the ability to transcend geographical and temporal barriers facilitate networking and the collective action of social groups (Bennett 2003). Whenever social groups seem to respond to threats, social identity may play an important role. For example, Internet users disparage scientific findings online if those findings threaten their social identity (Nauroth et al. 2015), and intergroup factors lead politically identified people to talk uncivilly in online news fora (Rains et al. 2017). Being rooted in sociological social psychology, social identity not only provides a perspective that explains micro behavior as embedded in social structures and social groups. It also provides testable hypotheses. To answer our research question, thus, social identity is optimal.

This study therefore draws on the social identity approach (Tajfel and Turner 1979, Turner, Hogg, Oakes et al. 1987, Branscombe, Ellemers, Spears et al. 1999) to explain behaviors between social groups. In addition, it draws on politicized social identity specifically (Simon and Klandermans 2001) to explain why these groups attack journalists. The combined approaches suggest five central conditions that increase threats to social identity and power: evaluative settings, the salience of politicized identities, disruptions of strong place identity, as well as evaluators that are powerful and belong to similar outgroups, i.e. other groups that are

hard to distinguish from the ingroup. This study applies them to journalists' characteristics to predict which characteristics elicit attacks (rather than focusing on the attacking audience members as an alternative feasible approach). It is hypothesized that journalists are more frequently attacked when their potential to threaten the power of social groups is greater. This potential is predicted to be greater if journalists are evaluative, publish on political topics, have a local focus, have more power, and belong to similar outgroups. These hypotheses are tested by an online survey of 530 journalists in Switzerland. The study makes three contributions. First, it proposes a theory-driven explanation of aggression against journalists in democracies that is oriented at sociological social psychology and empirically validates it. Second, combining social identity theory and politicized social identity to explain attacks on journalists stimulates theorization on the relation between the public and highly exposed professions in digital societies. Third, the "aggression divides" identified between attacked and unattacked journalists inform the development of counter-measures.

4.2. Social identity approach and politicized social identity

The social identity approach, comprising social identity theory and self-categorization theory, illuminates how social relationships affect behavior through social identity. Its underlying identity theory roots in Mead (1934) and symbolic interactionism (Stryker 1980), positing that the impact of society on behaviors is mediated by identities. Social identity theory (Tajfel and Turner 1979) argues that, dependent on the context, people slide from having personal identities, in which they see themselves and others as persons with individualized and unique character traits or abilities, to social identities in which they see themselves and others as depersonalized, typical representatives of social groups. Social identity is defined as individuals' awareness of their membership in a social group or groups, including the value and emotions attached to being members (Tajfel and Turner 1986). Socially identified people perceive their own group's members as more similar and their group as more different from outgroups (Hornsey 2008), and they conform to group norms, which are described as shared beliefs about what actions are appropriate in group-membership contexts (Bicchieri and Muldoon 2014). Social identities help individuals gain both intangible rewards such as belonging and being distinct and understood, and tangible rewards such as money. Further, they help groups achieve their interests and outcomes (Bicchieri and Muldoon 2014). By comparing their group with other groups, people thus strive to achieve and protect a positive social identity that depicts their group favorably and as positively distinct (Hornsey 2008).

If group members perceive their social identity to be threatened, they are motivated to discredit the threatening source. Members perceive threats if external sources such as outgroups and third parties decrease their group's value and esteem (Branscombe et al. 1999) and threaten group norms (Bicchieri and Muldoon 2014). In social identity theory, threats are not conceptualized as primarily objective, such as to jobs or wealth, but as subjective, such as feeling relatively deprived. In reality, of course, both types of threat are often linked. One strategy is to disparage the threatening source; group members perceive and treat the source of threat more negatively than their own group whenever opportunities for devaluation arise (Swann Jr and Schroeder 1995, Ellemers, Spears and Doosje 1997, Branscombe et al. 1999, Brewer 2007). For example, people verbally disparage the source (Scheepers, Spears, Doosje et al. 2003) and retaliate aggressively against it (Fischer, Haslam and Smith 2010). Accordingly, social identity threats correlate positively with outgroup disparagement (meta-analysis by Riek, Mania and Gaertner 2006) and cause it (e.g. Branscombe and Wann 1994).

While the social identity approach primarily explains bipolar behavior between two opposing groups, *politicized social identity*, a concept building on social identity theory, additionally and explicitly theorizes why third parties are addressed by group members. This perspective (Simon and Klandermans 2001, Klandermans 2014) acknowledges that groups try to establish and defend particular power structures to address their long-term and short-term grievances and achieve certain outcome distributions. The perspective emphasizes the societal embeddedness of such power struggles. It conceptualizes struggles as occurring not only between antagonistic groups. The general public and its representatives are influenced by each of the two antagonistic groups because as third parties they are sources of support and power. Third parties can also be authorities of the political system and news media. They are addressed and instrumentalized by group members who seek to enlist them as allies, pressure them to take sides, and influence and control them (Klandermans 2014). Third parties are addressed by politicized socially identified group members, who hold a form of social identity that underlies group members' explicit motivations to engage in power struggles as activists or within a political movement (Simon and Klandermans 2001, Van Zomeren, Postmes and Spears 2008). For example, news media can be addressed by a national conservative group that experiences its values being threatened by immigrants and that acts in its struggle for cultural hegemony (Simon and Klandermans 2001). Generally, any social identity may theoretically become politicized. However, this is more likely to occur in obvious, left-wing and right-wing ideologies and particular racial, ethnic, and religious groups. Occasionally, subgroups also politicize, such as feminists, and those identifying with specific issues, such as right-to-life on abortion (Huddy 2015). In contrast

to nonpoliticized social identities, such as merely identifying with one's gender, politicized social identities predict collective action more strongly (meta-analysis by Van Zomeren et al. 2008). This is because politicized social identity obliges people to participate in social activism more strongly (Stürmer and Simon 2004). Generally, any related actions can be conceptualized as collective as long as they aim at collective outcomes, even if they are perpetrated alone, as is commonly observed in digital environments (Postmes and Brunsting 2002).

4.3. Aggression against journalists

Applying politicized social identity theory to the phenomenon of aggression against journalists, this study assumes that journalists and audiences perceive and act in line with their social identities. This is for two reasons. First, if audiences read, listen, or view journalistic output in contexts where media reach large audiences via mass communication, audience members receive insufficient information about journalists to individualize them. Journalists and audience members do not form their relation through direct personal contact but through indirect analogue and digital channels such as print news articles and TV. How much information audiences receive about journalists varies between media channels. For example, TV moderators signal relatively rich verbal and nonverbal information, while authors of newspaper articles might reveal only their names. The overall relatively poor informational situation motivates audiences to perceive and treat journalists primarily in their journalistic role and to categorize them socially. Second, an intergroup context is obvious because journalistic content typically depersonalizes. Journalists publish more or less stereotype-compatible information on general matters and groups in society, including publications about individual group representatives. This content does not address each audience member personally but only indirectly through matters, groups, and representatives that audience members identify with. This motivates audiences to think and act as members of social groups instead of individuals, because potentially individualizing, personal contact with members of other groups seldom occurs in this context.

Journalists can be perceived by members of particular groups within audiences as threats to their social identity and power and consequently be aggressively disparaged. Journalists are not only part of the intergroup context but may also actively shape power relations between groups. This is because they embody the "media system", a potential third party in the wider societal arena, as suggested by Simon and Klandermans (2001). Journalists form groups' public images by setting agendas; they select and unearth particular stories and information and present these to the public. It can be assumed that group members wish to see their group portrayed positively,

distinctly, and favorably in the media landscape. Such a portrayal ultimately increases group power. If members perceive journalists as threatening the social identity and power of their group, they may seek to address these journalists. Such members are more likely to be politicized socially identified members than nonpoliticized members, as their inner obligation makes them more likely to take collective action (Van Zomeren et al. 2008) They may influence, pressurize, and try to control journalistic content on behalf of their group so as to improve the media image of the ingroup or worsen that of an outgroup. The aim of such action is to regain positive social identity and power. Aggression can occur reactively when published output triggers sudden grievances (e.g., animal rights activists being portrayed as militant), or proactively when journalists possess social characteristics that are perceived to threaten social hierarchies (e.g. journalists' nationality). Aggression is perpetrated by networks or by isolated individuals, physically such as by offending journalists face-to-face, and digitally, such as by posting offensive comments on social media.

The politicized social identity approach identifies various conditions that increase social identity threat. Journalists increase their threatening potential by being associated with these conditions and thus will be more frequently attacked. This study refers to potential because publishing under these conditions does not automatically lead to every journalist being attacked; this still depends on other factors such as portraying a certain group as positive or negative. However, it increases the chances of being attacked as group members with a pre-existing tendency to feel threatened are more easily 'mobilized'. Five conditions are considered here: evaluation, political topics through salience of politicized identities, local focus through threatening place-related identity processes, power, and similar outgroups through low intergroup distinctness.

Evaluation. People are more likely to feel threatened if they are evaluated. Social identity theory argues that group members evaluate by comparing groups with other groups (Tajfel and Turner 1986, Hogg 2000). Comparisons may result in positively distinct social identities but may also threaten social identities if groups come off badly. Comparative settings thus trigger group members' concerns about identity-contingent devaluation. Diverse studies experimentally induced identity-threatening information, for example by providing group members with negative feedback about their group's performing poorly or being disliked (Swann Jr and Schroeder 1995, De Hoog 2013). In one experiment, Branscombe, Spears, Ellemers et al. (2002) threatened the prestige of a group by telling them that outgroup members had evaluated them negatively. Group members consequently perceived lower collective self-esteem and allocated fewer resources to the threatening outgroup.

It follows that journalists that cultivate an evaluative publishing style are more frequently attacked because their critiques may threaten particular social groups. Some journalists do not solely publish information but evaluate the state of the world; they take positions on particular matters such as issues, ideas, values, and the achievements of groups and their representatives. These can be journalists who regularly publish opinion and editorial columns, comments, and leading articles. In addition, they can be regular journalists who belong to a journalistic culture that collectively values an evaluative style. Journalistic cultures have developed along language and cultural borders that not only divide media markets into separate segments but also influence journalistic roles and practices (Bonin, Dingerkus, Dubied et al. 2017). In evaluative, traditionally francophone, journalistic cultures, journalists tend to identify more strongly with a style that sets agendas, motivates citizens, and scrutinizes power than with one driven by attracting and satisfying audiences. For example, Bonin et al. (2017) found that evaluative styles are more predominant among francophone journalists in Switzerland than their other-language peers in the remaining language regions. These authors trace the evaluative style back to francophone minority identity, which has developed a sense of history distinct from the nationally dominant linguistic group – and a historically predominantly opinionated press.

Overall, evaluative journalists are more likely than those publishing relatively nonevaluative contents to bolster the power of some groups but weaken the power of others. They thus have a greater potential to threaten, and thus they attract aggression. This leads us to the following hypotheses:

> *H1.1: Journalists are more frequently attacked if they regularly publish opinionated articles than if they do not regularly publish opinionated articles.*
>
> *H1.2: Journalists are more frequently attacked if they are part of relatively evaluative journalistic cultures than if they are part of relatively audience-satisficing journalistic cultures.*

Political topics. The topics journalists select may make them threatening. According to self-categorization theory (Turner et al. 1987), group members feel threatened in situations where incoming information could threaten a social identity that is activated and relevant for the given situation. Members feel less threatened if irrelevant social identities and personal identities are activated (Brewer 2007). In one experiment, subjects exposed to country-threatening information about terror attacks felt more aggressive if national identity was activated rather than gender identity (Fischer et al. 2010). Generally, if particular social identities, whether permanently held or situationally induced, fit the situation where this identity is immediately

relevant, *salience* occurs (Turner et al. 1987, Huddy 2015). The fit increases the more the distinctions associated with a currently activated social identity maximize intragroup commonalities and intergroup differences (*comparative fit*) and confirm stereotypical expectations (*normative fit*). Salience can be induced experimentally, for example, by putting subjects in intergroup contexts and activating the appropriate social identities (Rip, Vallerand and Lafrenière 2012). Group members then conform more closely within the group, stereotype more, and respond to threats more aggressively.

Journalists publishing on political topics are liable to be more frequently attacked because these topics induce a fit between the situation and audience members' politicized social identities. Salience is here understood as an interaction between audiences' politicized or nonpoliticized social identities and the journalists' topics; political topics may make politicized social identities relevant, and non-political topics such as sports, culture, and technology make nonpoliticized identities relevant. If members with a politicized social identity are those that predominantly attack and they encounter political content, a comparative and normative fit is induced between this identity and its context. This is because politics is the societal topic where power is most proximately contested (Simon and Klandermans 2001); it is the naturally occurring context of politicized social identities. For example, journalists may publish on women from a political perspective (e.g. equal rights for women in the political section of the newspaper) or from a nonpolitical perspective (e.g. woman's make up in the lifestyle section). The former situation creates a fit with those members holding a politicized feminist social identity. They might feel threatened. Their inner obligation to act makes them easily mobilized and aggressive. The latter situation more likely creates a fit with those members holding a nonpoliticized gender social identity. They might also feel threatened. However, their weak inner obligation to collectively act make them less likely to attack. This leads us to the following hypothesis:

> *H2: Journalists that regularly publish on political topics are more frequently attacked than journalists who publish on other topics.*

Local focus. People also experience threat if place-related identity processes are disrupted. Social identities are not only represented by metaphorical places in society, such as those of housewives and politicians, but also by literal places (Dixon and Durrheim 2000). *Place identity* refers to how individuals' sense of identity is constituted by the physical and symbolic properties of locations (Proshansky, Fabian and Kaminoff 1983). Place identity is compatible with social identity theory (Bonaiuto, Breakwell and Cano 1996, Dixon and Durrheim 2000)

because it suggests that people identify with places to maintain self-esteem, continuity over time, self-efficacy, distinctiveness, and a sense of belonging. If processes are disrupted, people experience identity threat (Bonaiuto et al. 1996, Devine-Wright 2009). Local attachments seem particularly strong. The related research stream argues that individuals value and engage more in what is spatially and temporally immediate because it is less psychologically distant (Milfont 2010), more visible, tangible, and personally relevant, and it presents more opportunities to individuals for effective action (Lorenzoni, Nicholson-Cole and Whitmarsh 2007, Devine-Wright 2013). Social identity theory suggests that, when faced with a threat, locals cannot simply change their local identity as they could a global identity. In the absence of alternative, for example social mobility, strategies, they more likely disparage (Bonaiuto et al. 1996: 162). This explains widespread local-place-protective action, sometimes referred to as NIMBY (Not In My Back Yard) and the enthusiastic engagement in local environmental problems compared to passivity in tackling global climate change (Devine-Wright 2009: 426).

It follows that journalists that publish with a local focus are more frequently attacked because their potential to threaten strong place-related identity processes is higher. Empirical studies show that people expect journalistic reporting to satisfy place-identity-related needs. For example, Amsterdam city residents expect local media to serve social integration, local understanding, social cohesion, and belonging (Costera Meijer 2010). Journalists that engage with local topics are thus more likely to violate these expectations and pose a threat. Such strong local identities are likely to be found particularly in Switzerland. Swiss citizenry displays strong local and regional identities due to linguistic, religious, and socio-economic cleavages and strong traditions of local autonomy due to the highly decentralized political system (Hega 2001). Attacking journalists that locally publish is intended to protect this place identity. Compare this with journalists publishing on other-located (e.g. national, continental, global) or nonlocatable topics. Independently of how they publish, their potential to threaten and thus be aggressively targeted is lower. For example, even if globally identified audience members feel threatened by journalistic content, acting on the threat is less likely because it is psychologically more distant, less tangible, and personally less relevant, and global identities can be more easily stripped off. This leads us to the following hypothesis:

> *H3: Journalists that regularly publish with a local focus are more frequently attacked than journalists who publish on other-located or nonlocatable foci.*

Power. Journalists' position can also trigger aggression. Generally, the more powerful the third parties involved in a power struggle between groups are, the more likely group members are to

feel threatened. Although the politicized social identity approach does not establish explicit hypotheses on the extent of third parties' power to present perceived threat and attract disparagement, a number of researcher in this field mention its influence. For example, in proposing politicized social identity, Simon and Klandermans (2001: 324) suggest that group members seek "to win the support of third parties such as more powerful authorities" to turn their concerns into a matter of general interest. They seek "recognition of society or the larger community (e.g., the city, region, country, or European Union) as a more inclusive in-group membership". Just as some groups are more powerful than others (Simon and Klandermans 2001), so are third parties. It can be hypothesized that the more powerful third parties are, the better they serve groups' need to enlarge ingroup memberships but the better they can also lower groups' power by ignoring or even opposing them. For example, Bonaiuto et al. (1996) find that British subjects used strategies to cope with perceived threats to place identity, "especially if those were initially attributed by a powerful and disliked (…) institution" such as the EU (p. 160).

It follows that more powerful journalists are more frequently attacked because they are more able to shape power structures between groups. Journalists' power is greater if they work for media with large audiences. What those publishing companies and radio and television stations produce is read, seen, or heard by more people. Journalists are also more powerful if they are higher in the professional hierarchy. These journalists are not only opinion leaders but also have more influence about what is produced. Therefore, more powerful journalists have a greater threatening potential. These powerful actors should thus be more likely reached out to be influenced and controlled by groups. This leads us to the following hypotheses:

> *H4.1: The more powerful journalists are through working at media organizations with a larger reach, the more frequently they are attacked.*
>
> *H4.2: The more powerful journalists are through having a high rank in the professional hierarchy, the more frequently they are attacked.*

Low intergroup distinctiveness. Journalists' stable group characteristics may also trigger threat. Generally, group members feel more threatened if they perceive other groups to be too similar to their own group. The related *intergroup distinctiveness* describes "the perceived difference or dissimilarity between one's own group and another group on a relevant dimension of comparison" (Jetten, Spears and Postmes 2004: 862). Being different from other groups both legitimizes a group's existence and regulates the relation and interactions with other groups (Tajfel and Turner 1986, Jetten et al. 2004: 862). Group members thus feel threatened by other

groups that seem too similar to the ingroup in morals, norms, and values. Those might trigger, for example, cultural threats (Helbling 2011). Members then experience *distinctiveness threat*: their group's distinctiveness and uniqueness is threatened (Branscombe et al. 1999, Riek et al. 2006). In addition, they feel uncertain about their social identity, which otherwise provides relatively clear information to its members on how to think, feel, and behave (Hogg 2000: 227-228). While it is predominantly high identifiers that feel threatened by similar outgroups, low identifiers rely on more explicit differences to feel threatened such as by more dissimilar outgroups (even in this case, though, low identifiers are not sufficiently motivated to disparage). To restore distinctiveness and reduce uncertainty, group members will disparage the members of similar outgroups.

It follows that journalists belonging to outgroups that are very similar to the common audience ingroup are liable to be more frequently attacked than journalists belonging to the ingroup or more dissimilar outgroups. This is because they threaten the distinctiveness as perceived by the audience and make them uncertain about their identity. For Switzerland, this study considers journalists as similar if they have migration backgrounds from countries surrounding Switzerland. This is justified by studies distinguishing "culturally close", or similar, countries to Switzerland such as Italy or Germany from "culturally far" countries such as non-West European countries (Stolz 2000: 46). Although journalists with German migration backgrounds working in the German Swiss part or French journalists in the French Swiss part might be particularly threatening, it is assumed here that journalists from all the surrounding countries may threaten audiences in all Swiss parts. This is due to their common cultural closeness to Switzerland compared to that of culturally distant countries. Studies support this by showing that Swiss-Germans who perceive German immigrants as a cultural threat dislike Germans more (Helbling 2011) and that citizens in German Swiss cities can have negative attitudes against French and Italian people too (Stolz 2000). This leads us to the following hypothesis:

> *H5: Journalists that belong to similar outgroups are more frequently attacked than those belonging to the ingroup or more dissimilar outgroups.*

4.4. Method and Data

This study uses data from an online survey on journalists in Switzerland conducted between July and October 2017. The survey examined the frequency, form, and impact of aggression experienced by journalists in Switzerland. The formulation of some survey questions was inspired by two similar surveys in other countries by Preuss et al. (2017) and Löfgren Nilsson and Örnebring (2016). These survey questions were provided to the author on demand.

A common problem in research on journalists is that the journalistic population is not well defined, and the demarcation of the professional field is becoming increasingly unclear (Wyss and Keel 2010, Bonfadelli, Keel, Marr et al. 2011). In this study, the population is defined as freelance and employed journalists of print and online media (newspapers, magazines, news agencies), TV, and radio in the German-, French-, and Italian-speaking parts of Switzerland, excluding those who are retired and those working predominantly in advertising and public relations. This study relied on journalists responding to the survey to exclude themselves if they met either of these conditions. The population of Swiss journalists in 2017 is estimated to be approximately 10'500, based on figures in Bonfadelli et al. (2011). A two-step distribution channel maximized the reach of the survey. First, the survey, in the three national languages, was sent by email to all members of the four largest Swiss professional associations for journalists (*Impressum, Schweizer Syndikat Medienschaffender, Syndicom, Verband Schweizer Fachjournalisten*). This comprised 7877 journalists who were members of at least one of the four associations. This is the most common approach to surveying journalists in Switzerland because membership of an association is one of the preconditions for official registration as a journalist (Wyss and Keel 2010, Bonfadelli et al. 2011). The survey was also sent to all 6062 journalists registered at *renteria*, one of the largest Swiss journalist databases; to the author's knowledge, this is the first time this approach has been used. The two samples are assumed to strongly overlap, so the second step was a reminder for some association-registered journalists and an initial invitation for those not registered in any association.

Eventually, 637 questionnaires were completed, a response rate of approximately 8%. This is between response rates of 2% in a similar study on online aggression against journalists in Germany (Preuss et al. 2017) and of 12% (Oesch and Graf 2007) and 30% (Bonfadelli et al. 2011) in other-topic surveys on journalists in Switzerland. It is possible that attacked journalists were more likely to self-select into the survey, which would result in an overestimate of the prevalence of aggression. To minimize such a nonresponse bias, the survey explicitly invited all journalists, i.e. also those that have never been attacked, to participate. Social desirability bias was avoided by anonymized participation. The final sample can be considered representative for journalists in Switzerland. This is indicated by comparing our sample to the samples of two extensive studies on journalists in Switzerland by Oesch and Graf (2007) and Bonfadelli et al. (2011). Table 7 in the Appendix compares the socio-demographic composition of the three samples. They are similar in sex, age, employment position, education, media type, and language region. Slight differences may be ascribed to structural transformations in the

media landscape within the last 10 years. The present sample thus allows statistically meaningful conclusions to be drawn for all journalists in Switzerland.

Negative binomial regressions are applied in the statistical analyses to predict the frequency of an individual journalist being attacked (the first model includes the control variables and the second model adds the variables of interest). Negative binomial regressions suit the structure of the dependent variable, which is a count variable and has an over-dispersed negative binomial distribution: its variance is greater than its mean. A likelihood ratio test comparing the negative binomial regressions to a Poisson model strongly confirms the applicability of negative binomial regressions. The robustness tests applied are logit and Poisson regressions. They strongly confirm all results and therefore will not be further discussed.

4.4.1. Measurements

All variables refer to the situation within the 12 months prior to the survey, either continuously or at least most of the time. The questionnaire-based research design means that all the information acquired results from self-reports by individual journalists. The information below all arises from the final sample of 530 journalists. The case number of journalists is reduced as each journalist providing a "do not answer" on at least one variable is omitted.

4.4.2. Variables of interest

Frequency of being attacked is an ordinal variable indicating how often journalists were attacked. Journalists were asked how often they or their journalistic contents were targeted by "offenses, threats, aggressive, vulgar, pathologizing, or generalizing statements that are inappropriately disparaging, either directly such as through email, text message/phone, reader's letters, or face-to-face, or indirectly such as through public online channels including social networking sites, news commentary sections, or discussion fora". Some 43% of journalists were never attacked (= 0), 13% once (= 1), 15% once in 6 months (= 2), 16% once in 3 months (= 3), 8% once a month (= 4), 4% once a week (= 5), and 1% once daily (= 6). This variable is over-dispersed. Incidentally, aggression through digital channels seems to predominate: journalists were attacked at least once through private (47%) or public (42%) online channels, or through text message/phone, letter, or face-to-face (25%).

Evaluation by regularly publishing opinionated content is a dichotomous variable indicating evaluative publishing. Journalists were asked whether they "regularly published opinionated articles including journalistic columns, comments or leading articles". A majority, 54%, of journalists self-reported doing so (= 1), 46% not doing so (= 0).

Evaluation as part of evaluative journalistic culture is a dichotomous variable measuring whether journalists belong to the francophone culture in Switzerland or not. This was determined by allocating values of one to journalists who filled out the French survey (14%), and zero if not (i.e. if they filled out the Italian- or the German-speaking survey).

Publishing on political topics is a dichotomous variable. Journalists were asked on what topic(s) they regularly publish. Choosing from a list of thirteen topics, 52% of journalists self-reported publishing on political topics (= 1) while 48% did not (= 0).

Publishing with a local focus is a dichotomous variable that indicates whether journalists regularly published on local topics. Journalists were asked on what topic(s) they regularly publish. Choosing from the same list of thirteen topics as above, 48% of journalists self-reported publishing on local topics (= 1) while 52% did not (= 0).

Power by media reach of organization is an ordinal variable measuring the media reach of the organization journalists primarily worked for, commonly defined as the number of audience members. Journalists were asked how large the media's audience is: readers per issue for print, listeners or viewers per day for radio and TV, and unique users per day for online media. In all, 13% of journalists worked for an organization with a reach of 1 – 10,000 people, 21% of 10,001 – 50,000, 15% of 50,001 – 100,000, 35% of 100,001 – 500,000, 8% of 500,001 - 1 million, and 8% of 1 million - 5 million people.

Power by high rank in professional hierarchy is a dichotomous variable indicating whether journalists hold a position as leader of a section or team, for example as chief editor, or middle or senior management (35%; = 1), or have a lower rank as editor, freelancer, or trainee (65%, = 0).

Low intergroup distinctiveness by having migration background from surrounding countries is a dichotomous variable indicating whether journalists or one or more of their parents or grandparents immigrated from one or more of the countries surrounding Switzerland: France, Italy, Germany, and Austria (20%; = 1). The comparison group are journalists of more distinct migrant groups from nonsurrounding countries and journalists without migration background (80%; = 0).

4.4.3. Control variables

First, the study controls for socio-demographic information. The variables represent cues that may become the basis for social categorization, thus might trigger threat and consequently affect the frequency of being attacked.

Female is a dichotomous variable indicating whether journalists are female (35%; = 1) or male (65%; = 0). Research so far has provided inconsistent evidence on whether female journalists are either equally or more often attacked than males.

University degree is a dichotomous variable indicating whether journalists have a university degree or not. Overall, 80% have a tertiary degree while 20% finished either obligatory school or secondary school. Being highly educated may determine professional behaviors and thus influence aggression received. High education may also be a cue that attracts aggression, such as online comments discrediting scientific findings and scientists (Nauroth et al. 2015).

Age indicates journalists' age (⌀ = 46). As this variable correlates strongly with the years of journalistic experience, we omitted the latter in the analysis despite its potential relevance (Preuss et al. 2017).

Second, the study controls for *media type* indicating for which media type(s) journalists worked. Some 47% worked for subscription newspapers, 34% for (professional or news) magazines, 12% for radio, 11% for TV, 11% for online-only media, 6% for commuter/tabloid newspapers, and 4% for press agencies. Media type correlates with the frequency of being attacked, as also found by Preuss et al. (2017).

Third, the remaining *topics(s)* on which journalists regularly published are also controlled for. A majority, 60%, of journalists published on the topics of social affairs, culture, and entertainment, 41% on the economy and international affairs, 33% on science, technology, and the environment, 26% on crime and the judiciary, 18% on sports, and 8% on digital media and IT. These topic groupings emerged from factor analyses of the original topics. Topics correlate with the frequency of being attacked, as also found by Preuss et al. (2017) and Löfgren Nilsson and Örnebring (2016).

Fourth, the study also controls for other professional information.

Frequency of publishing is an ordinal variable indicating how often journalists published journalistic content. Just 1% of journalists published once in 6 months (= 1), 4% once in 3 months (= 2), 7% once a month (= 3), 16% several times a month (= 4), 10% once a week (= 5), 36% several times a week (= 6), 11% once daily (= 7), and 15% several times daily (= 8). Journalists publishing more often might be more frequently attacked as they provide more occasions to be perceived as threatening.

Frequency of social media activity is an ordinal variable indicating how actively journalists engaged in social media. They were asked how often they posted content or participated in

discussions using any semipublic or public social media profile such as Facebook, Twitter, and blogs. Some 38% were never active on social media (=0), 4% once (=1), 5% once in 6 months (=2), 4% once in 3 months (=3), 15% once a month (=4), 18% once a week (=5), 8% daily (=6), and 7% several times daily (=8). Social media activities may increase attacks as they negatively influence audience perceptions of journalists' professional roles (Lee 2015) while simultaneously making them more accessible.

Publicly accessible contact information is an ordinal variable indicating whether zero (14%), one (20%), two (27%), or all (39%) of the following personal data on journalists were publicly accessible: email address, private or office address, and mobile or office phone number. The more personal data that is publicly exposed, the more accessible journalists are. This increases the chances of being attacked also by less threatened, otherwise insufficiently motivated, group members.

Working full-time is a dichotomous variable indicating whether journalists worked full time as journalists (i.e. more than half of one's income coming from journalistic work or more than half of the work time used to produce journalistic content; 93%; = 1) or worked part time (7%; = 0). The level of employment may be associated with professional characteristics that influence being attacked.

For descriptive statistics and correlations of variables, see Table 8 in the Appendix.

4.5. Results

Table 4 reports the findings. The results in Model II confirm our hypotheses. Journalists are more frequently attacked if they are evaluative. For example, if they regularly publish opinionated contents, the frequency of attacks is 34% higher than for not regularly publishing opinionated contents (H1.1). Journalists who are part of a relatively evaluative journalistic culture report being attacked 47% more than those who are part of a relatively audience-serving journalistic culture (H1.2). Journalists are also more frequently attacked (frequency counts 45% higher) if they regularly publish on political topics than on other topics (H2). Further, journalists who regularly publish with a local focus report 35% more attacks than those publishing with other-located foci (e.g. national, continental, global) or nonlocatable foci (H3). Moreover, journalists are more frequently attacked when the reach of the media they primarily work for is larger (H4.1). Specifically, with every one unit increase in media reach of the organization, the frequency attack increases by 23%. Journalists are also more frequently attacked if they are a chief editor or middle or senior management (frequency counts increasing by 19%) than if they are an editor, freelancer, or volunteer (H4.2). However, the effect of H4.2 is only marginally

significant. Finally, journalists are more frequently attacked if they display low intergroup distinctiveness: if they have a migration background from countries surrounding Switzerland, compared to having no migration background or one from more distinct, nonsurrounding countries (H5). For this group, frequency of attacks is 29% higher.

The control variables indicate that journalists are more frequently attacked if they work for online-only media or TV and if they publish on the topic of crime and the judiciary. They are less frequently attacked if they work for press agencies and publish on social affairs, culture, and entertainment topics. Being attacked seems not to be significantly influenced by socio-demographic characteristics such as sex, education, or age. Likewise, working for radio or any of the magazines or newspapers does not impact the frequency of aggression. This also applies to the remaining topics: frequency of publishing, social media activity, the public accessibility of personal contact information, and working full-time.

Table 4. Effect of journalists' social identity threatening characteristics on their frequency of being attacked

	Model I			Model II		
	IRR	z		IRR	z	
Social identity threatening characteristics						
Evaluation by regularly publishing opinionated content				1.34	2.62	**
Evaluation as part of evaluative journalistic culture (French)				1.47	3.25	**
Publishing on political topics				1.45	3.24	**
Publishing with a local focus				1.35	2.78	**
Power by media reach of organization				1.23	4.87	***
Power by high rank in professional hierarchy				1.19	1.74	†
Low intergroup distinctiveness by having migration background from surrounding countries				1.29	2.24	*
Control variables						
Female	0.90	-0.95		0.93	-0.71	
University degree	1.06	0.44		1.02	0.16	
Age	1.00	-0.51		1.00	-0.99	
Media type: Working for …						
Press agency	0.37	-3.02	**	0.32	-3.52	***
Online-only media	1.27	1.59		1.59	3.31	**
Radio	0.83	-1.11		0.77	-1.64	
TV	1.48	2.30	*	1.34	1.87	†
(Professional, news) magazine	0.75	-1.86	†	0.86	-1.07	
Commuter/tabloid newspaper	1.21	0.94		0.91	-0.45	
Subscription paper	1.16	1.06		1.03	0.22	
Section: Regularly Publishing on …						
Social/Culture/Entertainment	0.93	-0.73		0.79	-2.40	*
Criminality/judiciary	1.37	2.73	**	1.26	2.07	*
Economy/international affairs	1.27	2.24	*	1.15	1.36	
Digital media/IT	1.03	0.17		1.11	0.59	
Sports	0.89	-0.83		0.88	-1.06	
Science/Environment	1.07	0.62		1.07	0.64	
Other professional characteristics						
Frequency of publishing	1.05	1.40		1.00	0.09	
Frequency of social media activity	1.03	1.22		1.00	-0.21	
Publicly accessible contact information	1.00	-0.07		0.99	-0.19	
Working full-time	1.40	1.42		1.07	0.28	
Constant	0.67	-0.90		0.41	-2.11	*
Number of observations		530			530	
LR chi2		74.08	***		147.55	***
df		20			29	
Pseudo R-square		4.18%			8.33%	

Legend: † p < .1, * p < .05, ** p < .01, *** p < .001

Models estimate a negative binomial regression. Effects are measured through Incidence Rate Ratio (IRR). The IRR represents the change in the dependent variable in terms of a percentage increase or decrease, with the precise percentage determined by the amount the IRR being either above or below 1.

4.6. Discussion

The results show that journalists are more frequently attacked by audiences if journalists are evaluative, publish on political topics, have a local focus, have more power, and if they belong to a similar outgroup. It is argued this is because these conditions increase journalists' potential to threaten the social identities of group members and the power of these groups. Evaluative journalists induce more comparative settings that potentially devalue groups than nonevaluative journalists. Journalists publishing on political topics make politicized identity salient. Journalists publishing with a local focus may disrupt strong place-related identity processes. More powerful journalists are more influential in shaping power structures between groups. Finally, journalists belonging to similar outgroups, compared to belonging to the ingroup or more dissimilar outgroups, threaten the distinctiveness and subjective certainty of the social identity shared by the average audience.

The first important contribution of this study is the theoretically and methodologically more sophisticated approach to aggression against journalists in democracies. While journalists have always been attacked, digital technology and social media seem to amplify it. This is highly likely as younger journalists enthusiastically engage in social media and older ones avoid it (Hedman and Djerf-Pierre. 2013); simultaneously, younger journalists report more perceived aggression than older ones (Preuss et al. 2017). This study engages with this emerging problem, uses theory oriented at sociological social psychology to embed it into a larger societal context, derives social identity-based hypotheses and validates them by multivariate analysis. This goes beyond the few existing, commonly atheoretical, descriptive, and case-based reports about this topic. Specifically, these results confirm certain conditions that make journalists to attract more aggression, as suggested by reports (Löfgren Nilsson and Örnebring 2016, Preuss et al. 2017). Results, though, disconfirm the absence of any effect of journalists' migration background in a former study (Preuss et al. 2017). The presence of this effect in the present study is explained by differentiating similar from dissimilar migration groups.

The second important contribution is the combination of the social identity approach (Tajfel and Turner 1986, Turner et al. 1987) with politicized social identity (Simon and Klandermans 2001) to theorize the relation between the public and highly exposed professions in digital societies. This approach seems appropriate, as the results reported here confirm central determinants of intergroup threat and disparagement: evaluation (e.g. Swann Jr and Schroeder 1995), salience of social identities (Turner et al. 1987), place identity (Proshansky et al. 1983, Devine-Wright 2009), the power of third parties (Simon and Klandermans 2001), and

intergroup distinctiveness (Branscombe et al. 1999, Jetten et al. 2004). Such a socially embedded perspective on aggression against public actors is in line with related and comprehensive social-norm explanations of collective outrages against public actors in social media (Rost et al. 2016). Both suggest that public actors today are easily caught in the fire of digital citizens fighting their struggles for norms, legitimacy, power, and identity on digital paths rather than on the street. The democratic potential of digital technology appears to be a strategic tool to aggressively exclude the unwanted and include the desired.

Ultimately, the findings suggest how to practically address aggression, or even counter it. First, they help to identify the most vulnerable journalistic groups. They point to aggression divides between frequently and rarely targeted journalistic groups. The findings suggest which journalist groups are most frequently targeted: evaluative, powerful journalists from similar outgroups publishing on local politics. Identification leads to the second point, intervention. Intervention aims to prevent members of such prototypical groups self-censoring, becoming silent, or quitting. Such understandable micro reactions may produce unintended negative macro outcomes. For example, if journalists individually accommodate the most aggressive social groups, the public media discourse may become biased in the long run. These results may thus inspire algorithms that estimate journalists' individual risk of being targeted based on their specific characteristics. This might help to calibrate the support provided to vulnerable groups early. Finally, the social identity perspective taken here may also inspire the restructuring of interactivity with audiences with the aim of reducing threats and disparagement. For example, interactions could be framed using individualizing rather than social category-based features, or superordinate social categories encompassing diverse groups instead of exclusive group categories (Brewer 2007). Both should reduce aggression, as they either activate audience members' personal identities or provide common group identities (Brewer 2007). These are alternatives to frequently used and drastic counterstrategies such as abolishing news commentary sections and ignoring audience members (Hedman and Djerf-Pierre 2013).

This study has several limitations that should be kept in mind when interpreting the results. First, the results can only be generalized to the population of journalists in Switzerland. However, the contextually sensitive social-identity perspective renders cross-national transferability conceivable. Second, the survey design does not allow causal conclusions to be drawn, but the inclusion of diverse control variables attempts to minimize the risk of potential confounding factors. Third, the interaction between journalists' characteristics and group members' perceptions of threat and politicized social identity was theoretically derived but not empirically tested. The present empirical approach is limited to directly relating journalists'

characteristics to the frequency of attacks on them. Inferring the operation of threat from being attacked instead of directly assessing it is a common approach in the literature on social identity threat (Branscombe et al. 1999).

The study opens important avenues for future research. First, future studies could test whether these hypotheses can be confirmed in other populations of journalists. Such efforts should incorporate socio-political structures, group hierarchies, and the status of journalists. Those structures may influence how audiences represent groups, perceive threats, and choose to respond. For example, while local place identity is presumably strong in Switzerland (Hega 2001), Hungarian citizens seem more strongly attached to their continent (Laczko 2005). Considering journalists' status, it is conceivable that in countries where few journalists are very popular, such as in the US, the aggression divide might be even deeper. While they are the winners-taking-it-all famewise, in terms of aggression received, they might simultaneously be the losers-getting-it-all. Second, the hypotheses could be tested to see whether they can predict aggression towards other third parties. For example, politicians have become similarly visible, accessible, and attacked by the public through social media such as their Facebook and Twitter profiles. Characteristics that make journalists vulnerable may well apply to politicians too. Third, alternative designs and samples could be used to analyze whether aggression against journalists is indeed driven by those audience members who are politically socially identified and feel threatened. This would acknowledge the heterogeneity within groups and confirm that aggression against journalists is most likely an interaction between the characteristics of journalists and of audience subgroups. This takes account of research suggesting that group members differ in how much they feel threatened, respond by disparaging, and choose hating and aggressive instead of peaceful collective action (Ellemers et al. 1997, Branscombe et al. 1999, Rip et al. 2012). An optimal approach would link survey data on audience members with behavioral data such as online hate comments against journalists.

To conclude, the interactivity between audiences and publicly exposed professions such as journalists is increasing in digital societies. This study has examined the dark side of this development, namely aggressive phenomena that occur in this process. The advocated explanatory perspective inspired by sociological social psychology hopes to initiate socially broader thinking while understanding, addressing, and potentially countering them.

5. References

Alonzo, Mei and Milam Aiken (2004). Flaming in electronic communication. *Decision Support Systems* 36(3): 205-213.

Álvarez-Benjumea, Amalia and Fabian Winter (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review* 34(3): 223–237.

Anderson, Ashley A., Dominique Brossard, Dietram A. Scheufele, et al. (2014). The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication* 19(3): 373-387.

Anderson, Ashley A., Sara K. Yeo, Dominique Brossard, et al. (2016). Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research* 30(1): 156–168.

Andreoni, James and Ragan Petrie (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics* 88(7-8): 1605–1623.

Antonetti, Paolo and Stan Maklan (2016). An extended model of moral outrage at corporate social irresponsibility. *Journal of Business Ethics* 135(3): 429-444.

Bandura, Albert (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review* 3(3): 193-209.

Barclay, Laurie J., David B. Whiteside and Karl Aquino (2014). To avenge or not to avenge? Exploring the interactive effects of moral identity and the negative reciprocity norm. *Journal of Business Ethics* 121(1): 15-28.

Bargh, John A., Katelyn Y. A. McKenna and Grainne M. Fitzsimons (2002). Can you see the real me? Activation and expression of the "true self" on the internet. *Journal of Social Issues* 58(1): 33-48.

Bauman, Sheri (2013). Cyberbullying: What does research tell us? *Theory Into Practice* 52(4): 249-256.

Bendor, Jonathan and Piotr Swistak (2001). The evolution of norms. *American Journal of Sociology* 106(6): 1493-1545.

Bennett, Winston (2003). Communicating global activism. *Information, Communication & Society* 6(2): 143-168.

Best, Henning and Clemens Kroneberg (2012). Die Low-cost-Hypothese: Theoretische Grundlagen und empirische Implikationen. *Kölner Zeitschrift für Soziologie* 64(3): 535–561.

Bicchieri, Cristina and Ryan Muldoon (2014). Social Norms. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.). Retrieved June 13, 2018, from: https://plato.stanford.edu/archives/spr2014/entries/social-norms/

Bishop, J. (2014). Representations of 'trolls' in mass media communication: a review of media-texts and moral panics relating to 'internet trolling'. *International Journal of Web Based Communities* 10(1): 7-24.

Bitektine, Alex and Patrick Haack (2015). The "macro" and the "micro" of legitimacy: Toward a multilevel theory of the legitimacy process. *Academy of Management Review* 40(1): 49-75.

Blau, Peter (1964). *Exchange and Power in Social Life*. New York: Wiley.

Blumer, Herbert (1969). *Symbolic Interaction: Perspective and Method*. Englewood Cliffs, NJ: Prentice-Hall.

Bonaiuto, Marino, Glynis M. Breakwell and Ignacio Cano (1996). Identity processes and environmental threat: The effects of nationalism and local identity upon perception of beach pollution. *Journal of Community & Applied Social Psychology* 6(3): 157-175.

Bonanno, Rina A. and Shelley Hymel (2013). Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying. *Journal of Youth and Adolescence* 42(5): 685-697.

Bonfadelli, Heinz, Guido Keel, Mirko Marr, et al. (2011). Journalists in Switzerland: Structures and attitudes. *Studies in Communication Sciences* 11(2): 7-26.

Bonin, Geneviève, Filip Dingerkus, Annik Dubied, et al. (2017). Quelle Différence? Language, culture and nationality as influences on francophone journalists' identity. *Journalism Studies* 18(5): 536-554.

Borah, Porismita (2013). Interactions of news frames and incivility in the political blogosphere: Examining perceptual outcomes. *Political Communication* 30(3): 456-473.

Borah, Porismita (2014). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research* 41(6): 809-827.

Branscombe, Nyla R., Naomi Ellemers, Russell Spears, et al. (1999). The context and content of social identity threat. In: N. Ellemers, R. Spears and B. Doosje (eds.). *Social Identity: Context, Commitment, Content* (35-58). Oxford, UK: Blackwell Science.

Branscombe, Nyla R., Russell Spears, Naomi Ellemers, et al. (2002). Intragroup and intergroup evaluation effects on group behavior. *Personality and Social Psychology Bulletin* 28(6): 744-753.

Branscombe, Nyla R. and Daniel L. Wann (1994). Collective self-esteem consequences of outgroup derogation when a valued social identity is on trial. *European Journal of Social Psychology* 24(6): 641-657.

Brauer, M. and P. Chekroun (2005). The relationship between perceived violation of social norms and social control: Situational factors influencing the reaction to deviance. *Journal of Applied Social Psychology* 35(7): 1519–1539.

Brewer, Marilynn B. (2007). The importance of being we: Human nature and intergroup relations. *American Psychologist* 62(8): 728-738.

Buckels, Erin E., Paul D. Trapnell and Delroy L. Paulhus (2014). Trolls just want to have fun. *Personality and Individual Differences* 67: 97-102.

Bundesministerium der Justiz und für Verbraucherschutz and juris GmbH (2016). Strafgesetzbuch (StGB): § 186 Üble Nachrede. Retrieved 13 June, 2018, from http://www.gesetze-im-internet.de/stgb/__186.html

Camerer, Colin F. (2011). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Cammaerts, B. and L. van Audenhove (2005). Online political debate, unbounded citizenship, and the problematic nature of a transnational public sphere. *Political Communication* 22(2): 147–162.

Castells, Manuel (2012). *Networks of Outrage and Hope: Social Movements in the Internet Age*. Cambridge, UK: Polity.

Cheng, Justin, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, et al. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *arXiv preprint arXiv:1702.01119*.

Cho, Daegon and Soodong Kim (2012). Empirical analysis of online anonymity and user behaviors: the impact of real name policy. 45th Hawaii International Conference on System Sciences (HICSS), Maui, HI 2012 4-7 Jan, IEEE

Coe, Kevin, Kate Kenski and Stephen A. Rains (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64(4): 658-679.

Coleman, James S. (1990). *Foundations of Social Theory*. Cambridge, MA: Belknap Press.

Connolly, Kate (6 Aug 2015). German TV presenter sparks debate and hatred with her support for refugees. Retrieved 13 June, 2018, from http://www.theguardian.com/world/2015/aug/06/german-tv-presenter-anja-reschke-sparks-debate-support-refugees

Conty, Laurence, Marisa Russo, Valerie Loehr, et al. (2010). The mere perception of eye contact increases arousal during a word-spelling task. *Social Neuroscience* 5(2): 171-186.

Cooper, Randolph B. and Russell Haines (2008). The Influence of workspace awareness on group intellective decision effectiveness. *European Journal of Information Systems* 17(6): 631-648.

Costello, Matthew, James Hawdon and Thomas N. Ratliff (2017). Confronting online extremism: The effect of self-help, collective efficacy, and guardianship on being a target for hate speech. *Social Science Computer Review* 35(5): 587-605.

Costera Meijer, Irene (2010). Democratizing journalism? Realizing the citizen's agenda for local news media. *Journalism Studies* 11(3): 327-342.

Crandall, Christian S. and Amy Eshleman (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin* 129(3): 414-446.

Crockett, M.J. (2017). Moral outrage in the digital age. *Nature Human Behaviour* 1(11): 769-771.

Dahrendorf, Ralf (1985/2010). *Homo Sociologicus*. Wiesbaden: Springer-Verlag.

David-Ferdon, Corinne and Marci F. Hertz (2007). Electronic media, violence, and adolescents: an emerging public health problem. *Journal of Adolescent Health* 41(6): 1-5.

De Hoog, Natascha (2013). Processing of social identity threats. *Social Psychology* 44(6): 361-372.

Deephouse, David L. and Mark Suchman (2008). Legitimacy in organizational institutionalism. *The Sage Handbook of Organizational Institutionalism* 49: 49-77.

Dennis, Alan R. (1996). Information exchange and use in group decision making: you can lead a group to information, but you can't make it think. *Small Group Research* 20(4): 433-457.

Dennis, Alan R. and Monica J. Garfield (2003). The adoption and use of GSS in project teams: toward more participative processes and outcomes. *Mis Quarterly* 27(2): 289-323.

Dennis, Kingsley (2008). Keeping a close watch–the rise of self-surveillance and the threat of digital exposure. *The Sociological Review* 56(3): 347-357.

Desai, Vinit M. (2011). Mass media and massive failures: Determining organizational efforts to defend field legitimacy following crises. *Academy of Management Journal* 54(2): 263-278.

DeSanctis, Gerardine and R. Brent Gallupe (1987). A foundation for the study of group decision support systems. *Management Science* 33(5): 589-609.

Devine-Wright, Patrick (2013). Think global, act local? The relevance of place attachments and place identities in a climate changed world. *Global Environmental Change* 23(1): 61-69.

Devine-Wright, Patrick (2009). Rethinking NIMBYism: The role of place attachment and place identity in explaining place-protective action. *Journal of Community & Applied Social Psychology* 19(6): 426-441.

Diekmann, Andreas and Peter Preisendörfer (1992). Persönliches Umweltverhalten. Diskrepanzen zwischen Anspruch und Wirklichkeit. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69: 226–251.

Diekmann, Andreas and Thomas Voss (2004). *Rational-Choice-Theorie in den Sozialwissenschaften*. München: Oldenbourg.

Diekmann, Andreas and Thomas Voss (2008). Soziale Normen und Reziprozitat. In: A. Diekmann, K. Eichner, P. Schmidt and T. Voss (eds.). *Rational Choice: Theoretische Analysen und empirische Resultate. Festschrift für Karl-Dieter Opp zum 70. Geburtstag* (83-100). Wiesbaden: VS Verlag.

DiMaggio, Paul, Eszter Hargittai, W. Russell Neuman, et al. (2001). Social implications of the Internet. *Annual Review of Sociology* 27(1): 307-336.

Dixon, John and Kevin Durrheim (2000). Displacing place-identity: a discursive approach to locating self and other. *British Journal of Social Psychology* 39(1): 27-44.

Dooley, Julian J., Jacek Pyzalski and Donna Cross (2009). Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. *Zeitschrift Fur Psychologie-Journal of Psychology* 217(4): 182-188.

Douglas Creed, W.E., Bryant A. Hudson, Gerardo A. Okhuysen, et al. (2014). Swimming in a sea of shame: Incorporating emotion into explanations of institutional reproduction and change. *Academy of Management Review* 39(3): 275-301.

Douglas, Karen M. and Craig McGarty (2001). Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British Journal of Social Psychology* 40: 399–416.

Durkheim, Emile (1957). *The Elementary Forms of the Religious Life, translated from French by Joseph Ward Swain*. London: G. Allen and UnWin.

Durkheim, Emile (1977). *Über die Teilung der sozialen Arbeit*. Frankfurt am Main Suhrkamp.

Einwiller, Sabine, Benno Viererbl and Sascha Himmelreich (2017). Journalists' coverage of online firestorms in German-language news media. *Journalism Practice* 11(9): 1178-1197.

Ellemers, Naomi, Russell Spears and Bertjan Doosje (1997). Sticking together or falling apart: In-group identification as a psychological determinant of group commitment versus individual mobility. *Journal of Personality and Social Psychology* 72(3): 617-626.

Ellickson, Robert C. (July 1999) The evolution of social norms: A perspective from the legal academy. *Social Science Research Network Paper Collection* (2-86).

Elster, Jon (1989). Social norms and economic theory. *Journal of Economic Perspectives* 3(4): 99-117.

Elster, Jon (2015). Social norms. In: J. Elster (eds.). *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge, United Kingdom: Cambridge University Press.

Erjavec, Karmen and Melita Poler Kovačič (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society* 15(6): 899-920.

Etter, Michael, Elanor Colleoni, Laura Illia, et al. (2017). Measuring organizational legitimacy in social media: Assessing citizens' judgments with sentiment analysis. *Business & Society* 57(1): 1-38.

Fairchild, Charles (2007). Building the authentic celebrity: The "idol" phenomenon in the attention economy. *Popular Music and Society* 30(3): 355-375.

Feather, N.T. and James W. Newton (1982). Values, expectations, and the prediction of social action: an expectancy-valence analysis. *Motivation and Emotion* 6(3): 217-244.

Fehr, E. and U. Fischbacher (2004). Social norms and human cooperation. *Trends in Cognitive Sciences* 8(4): 185-190.

Fehr, Ernst and Simon Gächter (2002). Altruistic punishment in humans. *Nature* 415(6868): 137-140.

Fehr, Ernst and Klaus Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114(3): 817-868.

Fischbacher, Urs, Ernst Fehr and Simon Gächter (2001). Are people conditionally cooperative? Evidence from public good experiments. *Economic Letters* 71(3): 397-404.

Fischer, Peter, S. Alexander Haslam and Laura Smith (2010). "If you wrong us, shall we not revenge?" Social identity salience moderates support for retaliation in response to collective threat. *Group Dynamics: Theory, Research, and Practice* 14(2): 143-150.

Franke, Nikolaus, Peter Keinz, Alfred Taudes, et al. (2017). The effectiveness of firm-controlled supporters to pacify online firestorms: A Case-based simulation of the "Playmobil" customize-it incident. In: U. Putro, M. Ichikawa and M. Siallagan (eds.).

Agent-Based Approaches in Economics and Social Complex Systems IX. Agent-Based Social Systems, vol 15 (153-164). Singapore: Springer.

Ganascia, Jean-Gabriel (2010). The generalized sousveillance society. *Social Science Information* 49(3): 489-507.

Ghose, Anindya and Panagiotis G. Ipeirotis (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering* 23(10): 1498-1512.

Gigerenzer, Gerd (2008). Why heuristics work. *Perspectives on Psychological Science* 3(1): 20-29.

Ginges, Jeremy and Scott Atran (2009). What motivates participation in violent political action. *Annals of the New York Academy of Sciences* 1167(1): 115-123.

Gopal, Abhijit and Pushkala Prasad (2000). Understanding GDSS in symbolic context: shifting the focus from technology to interaction. *Mis Quarterly* 24(3): 509-546.

Granovetter, Mark S. (1973). The strength of weak ties. *American Journal of Sociology* 78(6): 1360-1380.

Grimmer, Justin and Brandon M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267-297.

Gunther, Albert C., Daniel Bolt, Dina L. G. Borzekowski, et al. (2006). Presumed influence on peer norms: how mass media indirectly affect adolescent smoking. *Journal of Communication* 56(1): 52-68.

Guth, Werner and Stefan Napel (2006). Inequality aversion in a variety of games - an indirect evolutionary analysis. *Economic Journal* 116(524): 1037-1056.

Gutwin, Carl and Saul Greenberg (2002). A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)* 11(3-4): 411-446.

Haack, Patrick, Michael D. Pfarrer and Andreas Georg Scherer (2014). Legitimacy-as-feeling: How affect leads to vertical legitimacy spillovers in transnational governance. *Journal of Management Studies* 51(4): 634-666.

Haines, Russell, Jill Hough, Lan Cao, et al. (2014). Anonymity in computer-mediated communication: More contrarian ideas with less influence. *Group Decision and Negotiation* 23: 765–786.

Hale, Scott, Helen Margetts and Taha Yasseri (2013). Understanding the dynamics of internet-based collective action using Big Data: analysing the growth rates of internet-based

petitions. Annual Conference of the UK Political Studies Association, Cardiff, Wales (25-27 March 2013).

Hardaker, Claire (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research. Language, Behaviour, Culture* 6(2): 215–242.

Hardaker, Claire and Mark McGlashan (2016). "Real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics* 91: 80-93.

Harrington, C. Lee and Denise D. Bielby (1995). Where did you hear that - technology and the social organization of gossip. *Sociological Quarterly* 36(3): 607-628.

Haslam, Nick (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review* 10(3): 252-264.

Hauser, Florian, Julia Hautz, Katja Hutter, et al. (2017). Firestorms: Modeling conflict diffusion and management strategies in online communities. *The Journal of Strategic Information Systems* 26(4): 285-321.

Hawdon, James, Atte Oksanen and Pekka Räsänen (2017). Exposure to online hate in four nations: a cross-national consideration. *Deviant Behavior* 38(3): 254-266.

Hayne, Stephen C., Carol E. Pollard and Ronald E. Rice (2003). Identification of comment authorship in anonymous group support systems. *Journal of Management Information Systems* 20(1): 301-329.

Hedman, Ulrika and Monika Djerf-Pierre (2013). The social journalist: Embracing the social media life or creating a new digital divide? *Digital Journalism* 1(3): 368-385.

Hedström, Peter and Peter Bearman (2009). What is analytical sociology all about? An introductory essay. In: P. Hedström and P. Bearman (eds.). *The Oxford Handbook of Analytical Sociology* (3-24). Oxford, UK: Oxford University Press.

Hedström, Peter and Petri Ylikoski (2014). Analytical sociology and rational choice theory. In: G. Manzo (eds.). *Analytical Sociology: Norms, Actions and Networks* (57-70). West Sussex, UK: Wiley.

Hega, Gunther M. (2001). Regional identity, language and education policy in Switzerland. *Compare: A Journal of Comparative and International Education* 31(2): 205-227.

Helbling, Marc (2011). Why Swiss-Germans dislike Germans: Opposition to culturally similar and highly skilled immigrants. *European Societies* 13(1): 5-27.

Henrich, Joseph, Robert Boyd, Samuel Bowles, et al. (2001). In search of Homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review* 91(2): 73-78.

Hewett, Kelly, William Rand, Roland T. Rust, et al. (2016). Brand buzz in the echoverse. *Journal of Marketing* 80(3): 1-24.

Hmielowski, Jay D., Myiah J. Hutchens and Vincent J. Cicchirillo (2014). Living in an age of online incivility: Examining the conditional indirect effects of online discussion on political flaming. *Information, Communication & Society* 17(10): 1196-1211.

Hogg, Michael A. (2000). Subjective uncertainty reduction through self-categorization: A motivational theory of social identity processes. *European Review of Social Psychology* 11(1): 223-255.

Hogg, Michael A., Deborah J. Terry and Katherine M. White (1995). A tale of two theories: A critical comparison of identity theory with social identity theory. *Social Psychology Quarterly*: 255-269.

Hollenbaugh, Erin E. and Marcia K. Everett (2013). The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication* 18(3): 283–302.

Homans, George C. (1950). *The Human Group*. New York: Harpers.

Homans, George C. (1958). Social behavior as exchange. *American Journal of Sociology* 63(6): 597-606.

Hornsey, Matthew J. (2008). Social identity theory and self-categorization theory: A historical review. *Social and Personality Psychology Compass* 2(1): 204-222.

Hsueh, Mark, Kumar Yogeeswaran and Sanna Malinen (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research* 41(4): 557-576.

Huddy, Leonie (2015). Group identity and political cohesion. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*: 1-14.

Hughes, Megan and Johann Louw (2013). Playing games: The salience of social cues and group norms in eliciting aggressive behaviour. *South African Journal of Psychology* 43(2): 252–262.

Hutchens, Myiah J., Vincent J. Cicchirillo and Jay D. Hmielowski (2015). How could you think that?!?!: Understanding intentions to engage in political flaming. *New Media & Society* 17(8): 1201-1219.

Hutcherson, Cendri A. and James J. Gross (2011). The moral emotions: a social-functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology* 100(4): 719-737.

Jane, Emma A. (2015). Flaming? What flaming? The pitfalls and potentials of researching online hostility. *Ethics and Information Technology* 17(1): 65-87.

Jane, Emma A. (2015). "Your a Ugly, Whorish, Slut" Understanding E-bile. *Feminist Media Studies* 14(4): 531-546.

Jetten, Jolanda, Russell Spears and Tom Postmes (2004). Intergroup distinctiveness and differentiation: a meta-analytic integration. *Journal of Personality and Social Psychology* 86(6): 862–879.

Johnen, Marius, Marc Jungblut and Marc Ziegele (2017). The digital outcry: What incites participation behavior in an online firestorm? *New Media & Society*: 1-21.

Johnson, Norman A., Randolph B. Cooper and Wynne W. Chin (2009). Anger and flaming in computer-mediated negotiation among strangers. *Decision Support Systems* 46(3): 660-672.

Joinson, Adam N. (2007). Disinhibition and the internet. In: J. Gackenbach (eds.). *Psychology and the Internet (Second Edition)* (75-92). Burlington: Academic Press.

Jones, Joanne C., Gary Spraakman and Cristóbal Sánchez-Rodríguez (2014). What's in it for me? An examination of accounting students' likelihood to report faculty misconduct. *Journal of Business Ethics* 123(4): 645-667.

Jonsson, Stefan, Henrich R. Greve and Takako Fujiwara-Greve (2009). Undeserved loss: The spread of legitimacy loss to innocent organizations in response to reported corporate deviance. *Administrative Science Quarterly* 54(2): 195-228.

Jost, John T., Mahzarin R. Banaji and Brian A. Nosek (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology* 25(6): 881-919.

Kahneman, Daniel and Shane Frederick (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In: T. Gilovich, D. Griffin and D. Kahneman (eds.). *Heuristics and Biases: The Psychology of Intuitive Judgment* (49-81). New York: Cambridge University Press 49.

Kasumovic, Michael M. and Jeffrey H. Kuznekoff (2015). Insights into sexism: male status and performance moderates female-directed hostile and amicable behaviour. *PloS one* 10(7): e0131613.

Kayany, Joseph M. (1998). Contexts of uninhibited online behavior: Flaming in social newsgroups on Usenet. *Journal of the Association for Information Science and Technology* 49(12): 1135-1141.

Kiesler, Sara, Jane Seigel and Timothy W. McGuire (1984). Social psychological aspects of computer-mediated communication. *American Psychologist* 39(10): 1123-1134.

Kiesler, Sara, David Zubrow, Anne Marie Moses, et al. (1985). Affect in computer-mediated communication: an experiment in synchronous terminal-to-terminal discussion. *Human-Computer Interaction* 1(1): 77-104.

Kim, Su Jung, Rebecca Jen-Hui Wang, Ewa Maslowska, et al. (2016). "Understanding a fury in your words": The effects of posting and viewing electronic negative word-of-mouth on purchase behaviors. *Computers in Human Behavior* 54: 511-521.

Kincaid, Peter J., Robert P. Fishburne, Richard L. Rogers, et al. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel No. RBR-8-75, Naval Technical Training Command Millington TN Research Branch.

Klandermans, Bert (2003). Collective political action. In: D. O. Sears, L. Huddy and R. Jervis (eds.). *Oxford Handbook of Political Psychology* (670-709). Oxford: University Press.

Klandermans, Pieter G. (2014). Identity politics and politicized identities: Identity processes and the dynamics of protest. *Political Psychology* 35(1): 1-22.

Koban, Kevin, Jan-Philipp Stein, Valentin Eckhardt, et al. (2018). Quid pro quo in Web 2.0. Connecting personality traits and Facebook usage intensity to uncivil commenting intentions in public online discussions. *Computers in Human Behavior* 79: 9-18.

Kokkinos, Constantinos M., Nafsika Antoniadou and Angelos Markos (2014). Cyber-bullying: an investigation of the psychological profile of university student participants. *Journal of Applied Developmental Psychology* 35(3): 204-214.

Ksiazek, Thomas B. (2015). Civil interactivity: How news organizations' commenting policies explain civility and hostility in user comments. *Journal of Broadcasting & Electronic Media* 59(4): 556-573.

Ksiazek, Thomas B., Limor Peer and Andrew Zivic (2015). Discussing the news: Civility and hostility in user comments. *Digital Journalism* 3(6): 850-870.

Kuran, Timur (1997). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, Mass.: Harvard University Press.

Kwon, K. Hazel and Daegon Cho (2017). Swearing effects on citizen-to-citizen commenting online: A large-scale exploration of political versus nonpolitical online news sites. *Social Science Computer Review* 35(1): 84-102.

Kwon, K. Hazel and Anatoliy Gruzd (2017). Is aggression contagious online?A case of swearing on donald trump's campaign videos on youtube. Proceedings of the 50th

Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Hawaii, January 4-7 (2017), HICSS.

Laczko, Leslie S. (2005). National and local attachments in a changing world system: Evidence from an international survey. *International Review of Sociology—Revue Internationale de Sociologie* 15(3): 517-528.

Lawler, Edward J., Cecilia Ridgeway and Barry Markovsky (1993). Structural social psychology and the micro-macro problem. *Sociological Theory* 11(3): 268-290.

Lee, Eun-Ju (2007). Deindividuation effects on group polarization in computer-mediated communication: the role of group identification, public-self-awareness, and perceived argument quality. *Journal of Communication* 57(2): 385-403.

Lee, Jayeon (2015). The double-edged sword: The effects of journalists' social media activities on audience perceptions of journalists and their news products. *Journal of Computer-Mediated Communication* 20(3): 312-329.

Lee, Suk-Jae and James T. Tedeschi (1996). Effects of norms and norm-violations on inhibition and instigation of aggression. *Aggressive Behavior* 22(1): 17-25.

Li, Qing (2007). New bottle but old wine: a research of cyberbullying in schools. *Computers in Human Behavior* 23(4): 1777-1791.

Lindenmeier, Jörg, Christoph Schleer and Denise Pricl (2012). Consumer outrage: Emotional reactions to unethical corporate behavior. *Journal of Business Research* 65(9): 1364-1373.

Löfgren Nilsson, Monica and Henrik Örnebring (2016). Journalism under Threat: Intimidation and harassment of Swedish journalists. *Journalism Practice* 10(7): 880-890.

Lorenzoni, Irene, Sophie Nicholson-Cole and Lorraine Whitmarsh (2007). Barriers perceived to engaging with climate change among the UK public and their policy implications. *Global Environmental Change* 17(3-4): 445-459.

Lowry, Paul Benjamin, Jun Zhang, Chuang Wang, et al. (2016). Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research* 27(4): 962-986.

Mann, Steve and Joseph Ferenbock (2013). New media and the power politics of sousveillance in a surveillance-dominated world. *Surveillance & Society* 11(1/2): 18-34.

Marwick, Alice E. and Danah Boyd (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1): 114-133.

Mason, Kimberly L. (2008). Cyberbullying: a preliminary assessment for school personnel. *Psychology in the Schools* 45(4): 323-348.

Masullo Chen, Gina and Shuning Lu (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media* 61(1): 108-125.

Maurer, Andrea and Michael Schmid (2010). Erklärende Soziologie. In: A. Maurer and M. Schmid (eds.). *Erklärende Soziologie* (13-22). Wiesbaden: VS Verlag für Sozialwissenschaften.

McAdam, Doug (1986). Recruitment to high-risk activism - the case of Freedom Summer. *American Journal of Sociology* 92(1): 64-90.

McAdam, Doug and Ronelle Paulsen (1993). Specifying the relationship between social ties and activism. *American Journal of Sociology* 99(3): 640-667.

McLeod, Poppy L. (2000). Anonymity and consensus in computer-supported group decision making. *Research on Managing Groups and Teams* 3: 175-204.

McPherson, Miller, Lynn Smith-Lovin and James M Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1): 415-444.

Mead, George H. (1934). *Mind Self and Society: From the Perspective of a Social Behaviourist*. Chicago: Chicago University Press.

Mehari, Krista R., Albert D. Farrell and Anh-Thuy H. Le (2014). Cyberbullying among adolescents: Measures in search of a construct. *Psychology of Violence* 4(4): 399-415.

Meltzer, Kimberly (2015). Journalistic concern about uncivil political talk in digital news media: Responsibility, credibility, and academic influence. *The International Journal of Press/Politics* 20(1): 85-107.

Milfont, Taciano L. (2010). Global warming, climate change and human psychology. In: V. Corral-Verdugo, C. H. Garcia-Cadena and M. Frias-Armenta (eds.). *Psychological approaches to sustainability: Current trends in theory, research and practice* (19-42). New York: Nova Science Publisher.

Mishna, Faye, Michael Saini and Steven Solomon (2009). Ongoing and online: children and youth's perceptions of cyber bullying. *Children and Youth Services Review* 31: 1222–1228.

Moor, Peter J., Ard Heuvelman and Ria Verleur (2010). Flaming on youtube. *Computers in Human Behavior* 26(6): 1536-1546.

Moore, Michael J., Tadashi Nakano, Akihiro Enomoto, et al. (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior* 28(3): 861–867.

Myers, Daniel J. (2000). The diffusion of collective violence: Infectiousness, susceptibility, and mass media networks. *American Journal of Sociology* 106(1): 173-208.

Näsi, Matti, Pekka Räsänen, James Hawdon, et al. (2015). Exposure to online hate material and social trust among Finnish youth. *Information Technology & People* 28(3): 607-622.

Nauroth, Peter, Mario Gollwitzer, Jens Bender, et al. (2015). Social identity threat motivates science-discrediting online comments. *PloS one* 10(2): e0117476.

Nocentini, Annalaura, Juan Calmaestra, Anja Schultze-Krumbholz, et al. (2010). Cyberbullying: Labels, behaviours and definition in three European countries. *Australian Journal of Guidance and Counselling* 20(2): 129-142.

O'Sullivan, Patrick B. and Andrew J. Flanagin (2003). Reconceptualizing 'flaming'and other problematic messages. *New Media & Society* 5(1): 69-94.

Oehmer, Franziska (2011). Skandale im Spiegel der Zeit: Eine quantitative Inhaltsanalyse der Skandalberichterstattung im Nachrichtenmagazin Der Spiegel. In: K. Bulkow and C. Petersen (eds.). *Skandale. Strukturen und Strategien öffentlicher Aufmerksamkeitserzeugung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Oesch, Daniel and Roman Graf (2007). Löhne in den Medien 2006 [Wages in the media 2006]. Schweizerischer-Gewerkschaftverbund. Bern.

Olson, Mancur (1965). *The logic of collective action*. Cambridge, MA: Harvard University Press.

Opp, Karl-Dieter (2001). Why do people vote? The cognitive-illusion proposition and its test. *Kyklos* 54(2-3): 355-378.

Opp, Karl-Dieter (2002). When do norms emerge by human design and when by the unintended consequences of human action? The example of the no-smoking norm. *Rationality and Society* 14(2): 131-158.

Ostrom, Elinor (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives* 14(3): 137-158.

Pabian, Sara, Charlotte J.S. De Backer and Heidi Vandebosch (2015). Dark Triad personality traits and adolescent cyber-aggression. *Personality and Individual Differences* 75: 41-46.

Pace, Stefano, Bernardo Balboni and Giacomo Gistri (2017). The effects of social media on brand attitude and WOM during a brand crisis: Evidences from the Barilla case. *Journal of Marketing Communications* 23(2): 135-148.

Palazzo, Guido and Andreas G. Scherer (2006). Corporate legitimacy as deliberation: A communicative framework. *Journal of Business Ethics* 66(1): 71-88.

Parsons, Talcott (1960). *Structure and Processes in Modern Society*. New York: Free Press of Glencoe.

Parsons, Talcott (1964). *The Social System*. New York: Routledge & Paul.

Patchin, Justin W. and Sameer Hinduja (2006). Bullies move beyond the schoolyard: a preliminary look at cyberbullying. *Youth Violence and Juvenile Justice* 4(2): 148–169.

Peterson, Jillian and James Densley (2017). Cyber violence: What do we know and where do we go from here? *Aggression and Violent Behavior* 34: 193-200.

Pfeffer, Jürgen, Thomas Zorbach and Kathleen M. Carley (2014). Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications* 20(1-2): 117-128.

Pinsonneault, Alain and Nelson Heppel (1998). Anonymity in group support systems research: a new conceptualization, measures, and contingency framework. *Journal of Management Information Systems* 14(3): 89-108.

Poerksen, Bernhard (2018). *Die große Gereiztheit*. München: Carl Hanser Verlag.

Popitz, Heinrich (1980). *Die normative Konstruktion von Gesellschaft*. Tübingen: Mohr.

Posner, Richard A. and Eric B. Rasmussen (1999). Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics* 19(3): 369-382.

Postmes, Tom and Suzanne Brunsting (2002). Collective action in the age of the Internet: Mass communication and online mobilization. *Social Science Computer Review* 20(3): 290-301.

Prentice-Dunn, Steven and Ronald W. Rogers (1982). Effests of public and private seld-awareness on deindividuation and aggression. *Journal of Personality and Social Psychology* 43(3): 503-513.

Preuss, Madlen, Frederick Tetzlaff and Andreas Zick (2017). Publizieren wird zur Mutprobe. Studie zur Wahrnehmung von und Erfahrungen mit Angriffen unter JournalistInnen [Publishing becomes a test of courage: Study on perceptions and experiences of aggression among journalists]. Berlin, Mediendienst Integration.

Prinzing, Marlis (2015). Shitstorms. In: K. Imhof, R. Blum, H. Bonfadelli, O. Jarren and W. V. (eds.). *Demokratisierung durch Social Media? Mediensymposium* (153). Wiesbaden: Springer VS.

Proshansky, Harold M., Abbe K. Fabian and Robert Kaminoff (1983). Place-identity: Physical world socialization of the self. *Journal of environmental psychology* 3(1): 57-83.

Putnam, Robert D. (2000). Bowling alone. America's declining social capital. In: L. Crothers and C. Lockhart (eds.). *Culture and Politics* (223-234). New York: Palgrave Macmillan.

Quintana-Orts, Cirenia and Lourdes Rey (2018). Forgiveness and cyberbullying in adolescence: Does willingness to forgive help minimize the risk of becoming a cyberbully? *Computers in Human Behavior* 81: 209-214.

Rains, Stephen A., Kate Kenski, Kevin Coe, et al. (2017). Incivility and political identity on the internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication* 22(4): 163-178.

Raub, Werner, Vincent Buskens and Marcel A.L.M. Van Assen (2011). Micro-macro links and microfoundations in sociology. *The Journal of Mathematical Sociology* 35(1-3): 1-25.

Raub, Werner and Thomas Voss (2017). Micro-Macro Models in Sociology: Antecedents of Coleman's Diagram. In: B. Jann and W. Przepiorka (eds.). *Social Dilemmas, Institutions and the Evolution of Cooperation. Festschrift for Andreas Diekmann*. Berlin: De Gruyter.

Rauhut, Heiko and Ivar Krumpal (2008). Enforcement of social norms in low-cost and high-cost situations. *Zeitschrift für Soziologie* 37(5): 380-402.

Reicher, Stephen D., Russell Spears and Tom Postmes (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology* 6(1): 161-198.

Reinig, Bruce A. and Robert J. Mejias (2004). The effects of national culture and anonymity on flaming and criticalness in GSS-supported discussions. *Small Group Research* 35(6): 698–723.

Reiss, Steven (2004). Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of General Psychology* 8(3): 179-193.

Riek, Blake M., Eric W. Mania and Samuel L. Gaertner (2006). Intergroup threat and outgroup attitudes: A meta-analytic review. *Personality and Social Psychology Review* 10(4): 336-353.

Rindova, Violina P., Timothy G. Pollock and Mathew L.A. Hayward (2006). Celebrity firms: The social construction of market popularity. *Academy of Management Review* 31(1): 50-71.

Rip, Blanka, Robert J. Vallerand and Marc-André K. Lafrenière (2012). Passion for a cause, passion for a creed: On ideological passion, identity threat, and extremism. *Journal of Personality* 80(3): 573-602.

Rösner, Leonie, Stephan Winter and Nicole C. Krämer (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior* 58: 461-470.

Rost, Katja, Lea Stahel and Bruno S. Frey (2016). Digital social norm enforcement: Online firestorms in social media. *PloS one* 11(6): e0155923.

Rost, Katja and Antoinette Weibel (2013). CEO pay from a social norm perspective: the infringement and reestablishment of fairness norms. *Corporate Governance-an International Review* 21(4): 351-372.

Rowe, Ian (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society* 18(2): 121-138.

Ruiz, Carlos, David Domingo, Josep Lluís Micó, et al. (2011). Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. *The International Journal of Press/Politics* 16(4): 463-487.

Runions, Kevin C., Michal Bak and Thérèse Shaw (2017). Disentangling functions of online aggression: The Cyber-Aggression Typology Questionnaire (CATQ). *Aggressive Behavior* 43(1): 74-84.

Salek, Thomas A. (2015). Controversy Trending: The Rhetorical Form of Mia and Ronan Farrow's 2014 Online Firestorm Against# WoodyAllen. *Communication, Culture & Critique* 9(3): 477-494.

Salmivalli, Christina, Kirsti Lagerspetz, Kaj Bjorkqvist, et al. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior* 22(1): 1-15.

Sassenberg, Kai and Tom Postmes (2002). Cognitive and strategic processes in small groups: effects of anonymity of the self and anonymity of the group on social influence. *British Journal of Social Psychology* 41(3): 463-480.

Scheepers, Daan, Russell Spears, Bertjan Doosje, et al. (2003). Two functions of verbal intergroup discrimination: Identity and instrumental motives as a result of group identification and threat. *Personality and Social Psychology Bulletin* 29(5): 568-577.

Schelling, Thomas C. (2006). *Micromotives and Macrobehavior*. New York: WW Norton & Company.

Scherr, Albert (2013). Werte und Normen. In: A. Scherr (eds.). *Soziologische Basics* (271-278). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schmidtke, David S., Tobias Schröder, Arthur M. Jacobs, et al. (2014). ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods* 46(4): 1108-1118.

Silva, Leandro Araújo, Mainack Mondal, Denzil Correa, et al. (2016). Analyzing the Targets of Hate in Online Social Media. Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), Palo Alto, California, The AAAI Press.

Simon, Bernd and Bert Klandermans (2001). Politicized collective identity: A social psychological analysis. *American Psychologist* 56(4): 319-331.

Simons, Rachel N. (2015). Addressing Gender-Based Harassment in Social Media: A Call to Action. iConference 2015, California, USA, March 24-27.

Slonje, Robert and Peter K. Smith (2008). Cyberbullying: another main type of bullying? *Scandinavian Journal of Psychology* 49(2): 147-154.

Smith, Peter K., Jess Mahdavi, Manuel Carvalho, et al. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry* 49(4): 376-385.

Sontag, Lisa M., Katherine H. Clemans, Julia A. Graber, et al. (2011). Traditional and cyber aggressors and victims: A comparison of psychosocial characteristics. *Journal of Youth and Adolescence* 40(4): 392-404.

Stahel, Lea (2018). "Dirty journalists, all liars!" - A social identity explanation for why journalists are attacked by audiences. *Submitted to European Journal of Sociology*.

Stahel, Lea and Christopher Cohrs (2015). Socially shared representations of the Israel-Palestine conflict: An exploration among conflict outsiders. *Conflict & Communication Online* 14(1): 1-19.

Stahel, Lea and Katja Rost (2018). Legitimacy perceptions in online firestorms. *In the revisioning process to be resubmitted to the Journal of Business Ethics.*

Stanley, J. Woody and Christopher Weare (2004). The effects of internet use on political participation - evidence from an agency online discussion forum. *Administration & Society* 36(5): 503-527.

Steffgen, Georges, Andreas König, Jan Pfetsch, et al. (2011). Are cyberbullies less empathic? Adolescents' cyberbullying behavior and empathic responsiveness. *Cyberpsychology, Behavior, and Social Networking* 14(11): 643-648.

Stephens, Amanda N., Steven L. Trawley and Keis Ohtsuka (2016). Venting anger in cyberspace: Self-entitlement versus self-preservation in# roadrage tweets. *Transportation Research Part F: Traffic Psychology and Behaviour* 42: 400-410.

Stets, Jan E. and Peter J. Burke (2000). Identity theory and social identity theory. *Social Psychology Quarterly*: 224-237.

Sticca, Fabio, Sabrina Ruggieri, Françoise Alsaker, et al. (2013). Longitudinal risk factors for cyberbullying in adolescence. *Journal of Community & Applied Social Psychology* 23(1): 52-67.

Stieglitz, Stefan and Linh Dang-Xuan (2013). Emotions and information diffusion in social media - sentiment of microblogs and sharing behavior. *Journal of Management Information Systems* 29(4): 217-248.

Stolz, Jörg (2000). *Soziologie der Fremdenfeindlichkeit: Theoretische und empirische Analysen*. Frankfurt/New York: Campus Verlag.

Storbeck, Justin and Gerald L. Clore (2008). Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and Personality Psychology Compass* 2(5): 1824-1843.

Stroud, Natalie J., Joshua M. Scacco, Ashley Muddiman, et al. (2014). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-Mediated Communication* 20(2): 188-203.

Stryker, Sheldon (1980). *Symbolic Interactionism: A Social Structural Version*. Menlo Park, CA: Benjamin-Cummings Publishing Company.

Stryker, Sheldon and Peter J. Burke (2000). The past, present, and future of an identity theory. *Social Psychology Quarterly* 64(3): 284-297.

Stürmer, Stefan and Bernd Simon (2004). Collective action: Towards a dual-pathway model. *European Review of Social Psychology* 15(1): 59-99.

Suchman, Mark C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review* 20(3): 571-610.

Suddaby, Roy, Alex Bitektine and Patrick Haack (2017). Legitimacy. *Academy of Management Annals* 11(1): 451-478.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior* 7(3): 321-326.

Sullivan, Jonathan (2014). China's Weibo: Is faster different? *New Media & Society* 16(1): 24-37.

Sunstein, Cass R. (2005). Moral heuristics. *Behavioral and Brain Sciences* 28(4): 531-541.

Swann Jr, William B. and Daniel G. Schroeder (1995). The search for beauty and truth: A framework for understanding reactions to evaluations. *Personality and Social Psychology Bulletin* 21(12): 1307-1318.

Tajfel, H. and J.C. Turner (1986). The social identity theory of intergroup behavior. In: S. Worchel and W. G. Austin (eds.). *Psychology of Intergroup Relations* (7-24). Chicago: Nelson-Hall.

Tajfel, Henri and John C. Turner (1979). An integrative theory of intergroup conflict. In: S. Worchel and W. G. Austin (eds.). *The Social Psychology of Intergroup Relations* (33-47). Chicago: Nelson-Hall Publishers.

Tichy, Wolfgang (2013). Shitstorm - Eine (zivil) rechtliche Einführung. *ecolex - Zeitschrift für Wirtschaftsrecht* 5: 396-399.

Tofalvy, Tamas (2017). Online harassment of journalists in Hungary: Forms, coping mechanisms and consequences for press freedom. S. Griffen and J. Luque. https://ipi.media/, International Press Institute (IPI).

Tost, Leigh P. (2011). An integrative model of legitimacy judgments. *Academy of Management Review* 36(4): 686-710.

Treem, Jeffrey W., Stephanie L. Dailey, Casey S. Pierce, et al. (2016). What we are talking about when we talk about social media: A framework for study. *Sociology Compass* 10(9): 768-784.

Turner, John C., Michael A. Hogg, Penelope J. Oakes, et al. (1987). *Rediscovering the Social Group: A Self-Categorization Theory*. Cambridge, MA, US: Basil Blackwell.

Tutić, Andreas, Johannes Zschache and Thomas Voss (2015). Soziale Normen. In: N. Braun and N. Saam (eds.). *Handbuch Modellbildung und Simulation in den Sozialwissenschaften* (627-662). Wiesbaden: Springer VS.

Tuzovic, Sven (2010). Frequent (flier) frustration and the dark side of word-of-web: exploring online dysfunctional behavior in online feedback forums. *Journal of Services Marketing* 24(6): 446-457.

Valacich, Joseph S., Leonard M. Jessup, Alan R. Dennis, et al. (1992). A conceptual framework of anonymity in group support systems. *Group Decision and Negotiation* 1(3): 219-241.

van Stekelenburg, Jacquelien, Bert Klandermans and Wilco W. van Dijk (2011). Combining motivations and emotion: the motivational dynamics of protest participation. *Revista de Psicologia Social* 26(1): 91-104.

Van Zomeren, Martijn, Tom Postmes and Russell Spears (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin* 134(4): 504–535.

Van Zomeren, Martijn, Russell Spears, Agneta H. Fischer, et al. (2004). Put your money where your mouth is! Explaining collective action tendencies through group-based anger and group efficacy. *Journal of Personality and Social Psychology* 87(5): 649-664.

Vandebosch, Heidi and Katrien Van Cleemput (2008). Defining cyberbullying: a qualitative research into the perceptions of youngsters. *Cyberpsychology & Behavior* 11(4): 499-503.

Vandebosch, Heidi and Katrien Van Cleemput (2009). Cyberbullying among youngsters: profiles of bullies and victims. *New Media & Society* 11(8): 1349-1371.

Vargo, Chris J. and Toby Hopp (2017). Socioeconomic status, social capital, and partisan polarity as predictors of political incivility on Twitter: a congressional district-level analysis. *Social Science Computer Review* 35(1): 10-32.

Vergne, Jean-Philippe (2012). Stigmatized categories and public disapproval of organizations: A mixed-methods study of the global arms industry, 1996–2007. *Academy of Management Journal* 55(5): 1027-1052.

von Sikorski, Christian and Maria Hänelt (2016). Scandal 2.0: How valenced reader comments affect recipients' perception of scandalized individuals and the journalistic quality of online news. *Journalism & Mass Communication Quarterly* 93(3): 551-571.

Wagner, David G. and Joseph Berger (1985). Do sociological theories grow? *American Journal of Sociology* 90(4): 697-728.

Wahrman, Ralph (2010). Status, deviance, and sanctions: a critical review. *Small Group Research* 41(1): 91-105 (Reprinted from Small Group Behavior, 103: 203-223, 1972).

Weber, Max (1904). Die" Objektivität" sozialwissenschaftlicher und sozialpolitischer Erkenntnis. *Archiv für Sozialwissenschaft und Sozialpolitik* 19(1): 22-87.

Weber, Max (1978 [1922]). *Economy and Society*. Berkeley: University of California Press.

Weiss, Howard M., Kathleen Suckow and Russell Cropanzano (1999). Effects of justice conditions on discrete emotions. *Journal of Applied Psychology* 84(5): 786– 794.

Whelan, Glen, Jeremy Moon and Bettina Grant (2013). Corporations and citizenship arenas in the age of social media. *Journal of Business Ethics* 118(4): 777-790.

Wippler, Reinhard and Siegwart Lindenberg (1987). Collective phenomena and rational choice. In: J. C. Alexander, B. Giesen, R. Münch and N. J. Smelser (eds.). *The Micro-Macro Link* (135-152). Berkeley and Los Angeles: University of California Press.

Wolak, Janis, David Finkelhor, Kimberly J. Mitchell, et al. (2008). Online "predators" and their victims - myths, realities, and implications for prevention and treatment. *American Psychologist* 63(2): 111-128.

Workman, Michael (2012). Rash impulsivity, vengefulness, virtual-self and amplification of ethical relativism on cyber-smearing against corporations. *Computers in Human Behavior* 28(1): 217–225.

Wright, Michelle F. (2017). Adolescents' perceptions of popularity-motivated behaviors, characteristics, and relationships in cyberspace and cyber aggression: the role of gender. *Cyberpsychology, Behavior, and Social Networking* 20(6): 355-361.

Wyss, Vinzenz and Guido Keel (2010). Schweizer Journalismuskulturen im sprachregionalen Vergleich: Eine quantitative Längsschnittuntersuchung zu Strukturmerkmalen und Einstellungen [Swiss cultures of journalism in a regional-language comparison: A quantitative longitudinal study on structural characteristics and attitudes]. In: A. Hepp, M. Höhn and J. Wimmer (eds.). *Medienkultur im Wandel* (245-260). Konstanz: UVK.

Xu, Bo, Zhengchuan Xu and Dahui Li (2016). Internet aggression in online communities: a contemporary deterrence perspective. *Information Systems Journal* 26(6): 641-667.

Ybarra, Michele L. and Kimberly J. Mitchell (2004). Online aggressor/targets, aggressors, and targets: a comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry* 45(7): 1308-1316.

Ybarra, Michele L. and Kimberly J. Mitchell (2004). Youth engaging in online harassment: associations with caregiver-child relationships, Internet use, and personal characteristics. *Journal of Adolescence* 27(3): 319-336.

Ybarra, Michele L. and Kimberly J. Mitchell (2007). Prevalence and frequency of Internet harassment instigation: implications for adolescent health. *Journal of Adolescent Health* 41(2): 189-195.

Zuckerman, Ezra W. (1999). The categorical imperative: Securities analysts and the illegitimacy discount. *American Journal of Sociology* 104(5): 1398-1438.

6. Appendix

Document 1. Permission for using data of the platform openpetition.de

Permission for using and publishing all data of www.openpetition.de received from its proprietor

Jörg Mitzlaff, the proprietor of the social media website www.openpetition.de gives his consent to the authors of the paper *Digital Social Norm Enforcement: Online Firestorms in Social Media* (Katja Rost, Lea Stahel, Bruno S. Frey) and to PLOS to use and publish all received data originating on www.openpetition.de under the condition to always name openPetition as the source of the data.

Berlin, 04.01.2016

Jörg Mitzlaff

CEO

openPetition gGmbH
Haus der Demokratie
Greifswalder Str. 4
10405 Berlin
www.openpetition.de

Table 5. Descriptive statistics and bivariate correlations

ID	Variable	Obs	Mean	Std.Dev.	Min	Max	1	2	3	4	5
1	Amount of online aggression (log)	566052	.19	.40	.00	2.77					
2	Anonymity	566053	.30	.46	.00	1.00	-.01				
3	Intrinsic motivation (log)	566053	.28	.51	.00	3.09	.03	-.02			
4	Status of the accused (log)	554782	2.06	.32	.69	2.40	.04	.06	.02		
5	Controversy of accusation	554782	.58	.22	.00	.75	.04	.05	.05	.13	
6	Accused is a natural person (vs. legal entity)	554782	.04	.19	.00	1.00	.03	-.01	.02	-.04	.03
7	Anonymity of social environment (log)	542900	7.88	2.61	.69	12.63	-.03	.02	.01	.00	-.02
8	Accusation is connected to a scandal	554780	.23	.42	.00	1.00	.03	.01	.08	.20	.00
9	Length of comment in words	566053	20.10	12.98	1.00	57.00	.11	-.01	.21	-.05	.01
10	Number of protesters (log)	554781	4.24	1.89	.00	7.98	.05	.05	.00	.19	.15
11	Time of comment submission after petition opening	566053	50343.15	61289.32	.00	628470.00	.05	.07	.07	.27	.21
12	Scope of protest	554782	.70	.28	.00	1.00	.09	.05	.02	.19	.18
13	Success of the petition	554782	.16	.36	.00	1.00	.05	.05	-.05	.19	.16
14	Motives: Income/minimization of costs	554782	.39	.49	.00	1.00	.01	.06	-.03	.14	.20
15	Motive: Security/social order/traditional values	554782	.09	.29	.00	1.00	.00	.00	-.05	.00	.01
16	Motive: Independence/self-determination	554782	.26	.44	.00	1.00	.03	.05	.04	.26	.22
17	Motive: Increasing life quality and competence	554782	.40	.49	.00	1.00	-.11	-.05	-.05	-.19	-.14
18	Topic: Art/culture/education	554782	.23	.42	.00	1.00	-.07	-.02	-.08	-.16	-.07
19	Topic: Economics	554782	.12	.32	.00	1.00	.01	-.01	-.08	-.10	.04
20	Topic: Politics	554782	.08	.27	.00	1.00	-.04	.02	.06	.04	.05
21	Topic: Media	554782	.17	.38	.00	1.00	.11	.12	-.04	.36	.21
22	Topic: Environmental and animal welfare	554782	.08	.27	.00	1.00	.06	-.06	.02	-.10	-.01

ID	Variable	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
7	Anonymity of social environment (log)	-.02															
8	Accusation is connected to a scandal	.06	.02														
9	Length of comment in words	.00	.00	-.03													
10	Number of protesters (log)	-.07	-.02	.20	-.02												
11	Time of comment submission after petition opening	-.07	.01	.22	-.03	-.02											
12	Scope of protest	.01	-.19	.15	-.04	.21	.18										
13	Success of the petition	-.08	.00	-.13	-.04	.24	.15	.22									
14	Motives: Income/minimization of costs	-.05	.01	-.07	.03	.29	.20	.04	.20								
15	Motive: Security/social order/traditional values	.00	-.02	.05	-.01	-.11	.02	.02	.05	-.16							
16	Motive: Independence/self-determination	-.04	.01	.06	-.04	.19	.21	.19	.36	.01	-.09						
17	Motive: Increasing life quality and competence	-.05	.10	-.05	.03	-.20	-.20	-.29	-.14	-.20	-.07	-.14					
18	Topic: Art/culture/education	-.08	.08	-.10	.08	.05	-.21	-.16	-.17	.18	-.13	-.22	.29				
19	Topic: Economics	-.04	-.10	.00	-.01	-.12	-.02	.07	.01	-.15	.22	.03	-.08	-.19			
20	Topic: Politics	.09	.00	.17	-.04	.06	-.01	-.03	-.07	-.02	-.05	.11	-.09	-.16	-.10		
21	Topic: Media	-.02	.02	.07	-.06	.25	.42	.31	.46	.38	-.09	.32	-.27	-.24	-.16	-.13	
22	Topic: Environmental and animal welfare	.03	-.09	-.12	-.01	-.15	-.07	.13	-.05	-.23	-.07	-.17	-.14	-.16	-.11	-.09	-.13

Table 6. Descriptive statistics and bivariate correlations

	Variables	Obs	Mean	Std. Dev.	Min	Max	1	2	3	4	5	6	7	8
1	Aggressive punishment	45,997	0.19	0.52	0.00	7.00								
2	Emotional shaming	45,985	0.29	0.52	0.00	4.00	0.05							
3	Sober disapproval	45,997	0.64	0.48	0.00	1.00	-0.48	-0.75						
4	Illegitimate character	45,997	0.06	0.26	0.00	4.00	0.10	0.03	-0.06					
5	Illegitimate procedures	45,997	0.26	0.60	0.00	8.00	0.04	0.04	-0.04	0.13				
6	Illegitimate structures	45,997	0.03	0.17	0.00	4.00	0.02	-0.02	0.01	0.06	0.06			
7	Illegitimate outcomes	45,997	0.10	0.38	0.00	6.00	0.03	0.03	-0.04	0.12	0.13	0.06		
8	Sophistication of language	45,985	7.76	6.22	0.00	55.50	-0.06	-0.09	0.13	0.10	0.19	0.09	0.07	
9	Arousal through spelling mistakes	45,985	1.07	2.15	0.00	55.00	0.11	0.07	-0.07	0.12	0.14	0.04	0.07	0.27
10	Arousal level of words	45,985	4.26	0.67	2.30	8.10	0.03	-0.01	-0.01	0.04	0.06	0.01	-0.01	0.11
11	Anonymity	45,997	0.48	0.50	0.00	1.00	-0.04	-0.01	0.03	-0.01	0.00	-0.01	-0.01	0.05
12	Self-reported economic dependency	45,997	0.05	0.21	0.00	1.00	-0.04	-0.01	0.04	-0.03	-0.02	-0.01	-0.02	0.09
13	Urbanity of residence	44,173	0.46	0.95	0.00	3.07	-0.02	-0.03	0.03	-0.01	0.00	0.01	-0.01	-0.01
14	Negative media coverage	45,997	3.32	4.21	0.00	24.00	0.01	0.01	-0.01	0.01	0.00	0.00	-0.01	-0.02
15	Balanced media coverage	45,997	1.76	2.51	0.00	14.00	-0.01	0.01	0.00	0.00	0.00	0.00	-0.02	-0.01
16	Previous comments (total)	45,985	22.99	13.27	0.00	45.99	0.00	0.01	0.00	-0.01	-0.01	-0.02	-0.03	-0.01
17	Previous aggressive commenting	45,985	1.88	1.66	0.00	15.00	0.01	0.00	-0.01	0.00	0.01	0.00	0.00	-0.01
18	Number of words	45,985	0.22	0.26	0.01	4.75	0.06	0.10	-0.06	0.21	0.30	0.12	0.17	0.49

Variables	9	10	11	12	13	14	15	16	17
10 Arousal level of words	0.11								
11 Anonymity	-0.02	0.00							
12 Self-reported economic dependency	0.02	0.00	0.00						
13 Urbanity of residence	0.02	0.03	0.00	0.00					
14 Negative media coverage	0.00	0.01	0.00	-0.02	0.00				
15 Balanced media coverage	0.00	0.01	0.00	-0.02	0.04	0.36			
16 Previous comments (total)	0.00	0.02	-0.01	-0.05	0.05	0.22	0.24		
17 Previous aggressive commenting	0.00	0.00	0.00	-0.01	-0.01	0.04	-0.03	0.00	
18 Number of words	0.48	0.20	0.02	0.09	0.00	0.00	0.00	0.02	0.00

Table 7. Comparisons of socio-demographic information of the present journalist sample with those of former surveys (%)

Authors		Oesch and Graf	Bonfadelli et al.	Present study
Survey		Online	Online / printed	Online
Year of survey		2006	2006/2007/2008	2017
Estimated population (N)		10'000	10'500	10'500
Sample (N)		1157	2509	530
Region	German speaking	72	72	80
	French speaking	24	26	14
	Italian speaking	4	2	6
	In total (%)	100	100	100
Gender	Female	35	35	35
	Male	65	65	65
Age	Mean Age	(-)	43	46
Education	Compulsory	2	(-)	1
	Secondary	36	(-)	19
	Tertiary studies	62	56	74
	Doctorate			6
	In total (%)	100	56	100
Media type*	Television	28	31	11
	Radio			12
	Tabloid / free commuter newspaper	48	3	6
	(Sunday/weekly /daily) subscription newspapers			47
	(News/specialist) magazines	17	57	34
	News agency / media service	4	4	4
	Exclusively online media	3	3	11
	Other	(-)	2	(-)
	In total (%)	100	100	(multiple response)
Employment position*	Freelance	9	18	8
	Trainee	2	11	1
	Editor	68	49	65
	Team leader, chief editor, management	21	40	35
	In total (%)	100	(multiple response)	(multiple response)

If studies collected and reported information on several categories in a single, broader category, the relevant categories are framed in black.
*The comparability of the values in these categories is limited due to differing collection method used (single versus multiple response allowed)

Table 8. Descriptives and correlations

	Variables	Obs	Mean	Std. Dev.	Min	Max	1	2	3	4	5	6	7	8	9	10	11
1	Frequency of being aggressed	530	1.48	1.60	0	6											
2	Evaluation by regularly publishing opinionated content	530	0.54	0.50	0	1	0.12										
3	Evaluation by being part of evaluative journalistic culture (French)	530	0.14	0.35	0	1	0.19	0.02									
4	Salience of politicized identity by publishing on politics	530	0.52	0.50	0	1	0.33	0.10	0.07								
5	Potential to disrupt local identity by publishing on local issues	530	0.48	0.50	0	1	0.21	0.03	0.04	0.30							
6	Power by media reach of organization	530	3.26	1.43	1	6	0.19	-0.23	0.06	0.16	-0.09						
7	Power by high rank in professional hierarchy	530	0.35	0.48	0	1	0.10	0.18	0.00	0.10	0.01	-0.04					
8	Low intergroup distinctiveness by having migration background from surrounding countries	530	0.20	0.40	0	1	0.08	0.07	0.05	-0.05	0.01	-0.02	0.01				
9	Female	530	0.35	0.48	0	1	-0.10	-0.15	0.00	-0.21	-0.09	0.01	-0.10	0.16			
10	University degree	530	0.79	0.40	0	1	0.01	-0.04	0.03	0.01	-0.09	0.00	0.01	-0.04	0.09		
11	Age	530	45.90	11.39	21	74	-0.09	0.17	0.04	-0.16	-0.22	-0.16	0.20	0.02	-0.13	-0.14	
12	Press agency	530	0.04	0.20	0	1	-0.10	-0.21	0.03	0.07	-0.05	0.23	-0.05	-0.03	-0.03	0.06	0.04
13	Online-only media	530	0.11	0.32	0	1	0.09	-0.06	-0.04	0.00	-0.08	-0.02	0.01	0.02	0.00	0.03	0.04
14	Radio	530	0.12	0.33	0	1	0.00	-0.18	-0.05	0.09	0.02	0.20	0.04	-0.04	-0.06	-0.01	-0.06
15	TV	530	0.11	0.32	0	1	0.13	-0.22	0.05	0.13	0.07	0.14	-0.04	-0.07	0.01	0.08	-0.09
16	(Professional, news) magazine	530	0.34	0.47	0	1	-0.24	0.02	-0.08	-0.34	-0.27	-0.31	0.10	0.07	0.14	0.10	0.18
17	Commuter/tabloid newspaper	530	0.06	0.24	0	1	0.09	-0.13	0.08	0.02	0.01	0.30	0.05	-0.05	0.00	-0.03	-0.05
18	Subscription paper	530	0.47	0.50	0	1	0.13	0.35	0.04	0.14	0.25	-0.09	-0.05	0.03	-0.13	-0.03	-0.05
19	Social/Culture/Entertainment	530	0.60	0.49	0	1	-0.04	-0.07	0.04	0.13	0.07	0.02	0.08	0.07	0.13	-0.02	0.04
20	Criminality/judiciary	530	0.26	0.44	0	1	0.21	-0.13	-0.01	0.44	0.29	0.16	-0.03	-0.06	-0.12	-0.06	-0.17
21	Economy/international affairs	530	0.41	0.49	0	1	0.14	0.03	0.01	0.31	-0.03	0.07	0.08	-0.02	-0.18	0.06	-0.05
22	Digital media/IT	530	0.08	0.27	0	1	0.05	0.00	0.03	0.08	0.04	0.00	0.01	0.02	-0.08	-0.01	-0.03
23	Sports	530	0.18	0.38	0	1	0.03	-0.04	-0.02	0.09	0.17	0.06	0.00	-0.04	-0.15	-0.06	-0.11
24	Science/Environment	530	0.33	0.47	0	1	0.02	-0.06	0.05	0.07	-0.03	-0.02	0.04	0.04	0.00	0.05	0.04
25	Frequency of publishing	530	5.57	1.69	1	8	0.17	0.02	0.12	0.27	0.30	0.29	-0.15	-0.03	-0.16	-0.05	-0.22
26	Frequency of social media activity	530	2.78	2.53	0	7	0.13	0.12	0.10	0.09	0.03	0.07	0.02	0.08	-0.05	-0.05	-0.12
27	Publicly accessible contact information	530	1.91	1.08	0	3	-0.01	0.09	-0.03	-0.02	0.03	-0.18	0.19	0.05	-0.07	-0.03	0.20
28	Working full-time	530	0.93	0.25	0	1	0.11	0.06	0.09	0.15	0.05	0.18	0.05	0.04	-0.01	-0.02	-0.09

	Variables	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
13	Online-only media	-0.04															
14	Radio	-0.05	0.01														
15	TV	0.05	-0.01	0.20													
16	(Professional, news) magazine	-0.03	0.00	-0.16	-0.21												
17	Commuter/tabloid newspaper	-0.01	0.01	-0.09	-0.09	-0.10											
18	Subscription paper	-0.18	-0.06	-0.21	-0.26	-0.45	-0.06										
19	Social/Culture/Entertainment	0.07	0.11	0.05	0.06	0.13	-0.02	-0.19									
20	Criminality/judiciary	0.13	-0.05	0.06	0.13	-0.29	0.08	0.04	0.10								
21	Economy/international affairs	0.08	0.09	0.15	0.08	-0.10	0.00	-0.07	0.05	0.12							
22	Digital media/IT	0.08	0.07	0.13	-0.01	0.02	0.02	-0.10	0.09	0.07	0.22						
23	Sports	0.03	0.02	0.12	0.08	-0.13	-0.01	0.01	0.04	0.07	0.06	0.12					
24	Science/Environment	0.14	0.05	0.04	0.13	0.11	-0.06	-0.21	0.18	0.11	0.11	0.17	0.09				
25	Frequency of publishing	0.15	0.09	0.16	0.02	-0.42	0.16	0.18	-0.07	0.27	0.09	0.16	0.21	-0.05			
26	Frequency of social media activity	-0.05	0.17	0.03	-0.02	-0.15	-0.01	0.14	0.09	0.03	0.06	0.22	0.04	0.02	0.14		
27	Publicly accessible contact information	-0.11	0.04	0.03	-0.08	0.13	0.01	0.02	0.00	-0.06	0.05	0.01	-0.04	0.04	-0.09	-0.08	
28	Working full-time	0.06	-0.07	0.01	0.05	-0.15	0.00	0.11	-0.03	0.12	0.07	0.05	-0.02	-0.04	0.26	0.04	-0.01

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Philosophische Fakultät
Studiendekanat
Rämistrasse 69
CH-8001 Zürich
www.phil.uzh.ch

Erklärung

Hiermit erkläre ich, dass die Dissertation von mir selbst ohne unerlaubte Beihilfe verfasst worden ist und diese Dissertation noch an keiner anderen Fakultät eingereicht wurde.

Ort und Datum: Zürich, 26.06.2018

Unterschrift: